Soutenance Projet OC P2 DS: Analysez les systèmes éducatifs

Candidat: David CAPELLE
Mentor: Nicolas MICHEL

Evaluateur: Mohammed SEDKI
Formation 100% Pôle Emploi

20/05/2021

Plan de l’analyse pré-exploratoire des données

- Rappel de la problématique
- Description et simplification du jeu de données
- Sélection d'indicateurs avec plusieurs étapes de filtres
- Analyse de la distribution des données pour les indicateurs retenus
- Calcul d’un score d’attractivité par pays retenus
- Conclusion sur la pertinence du jeu de données

Description du jeu de données – 5 datasets (1/2)

- Dataset "country" (EdsStatscountry.csv)
  - 241 lignes, 32 variables, liste des 241 pays avec des informations macro-économiques sur l'économie des pays
- Dataset "countryseries" (EdStatsCountry-Series.csv)
  - 613 lignes, 4 variables, informations sur la source des données (variable « Description ») des indicateurs par pays présents dans le dataset **country**.
- Dataset "footnote" (EdStatsCountry-Series.csv)
  - 643638 lignes, 5 variables, information (variable « Description ») sur l'année d'origine des indicateurs présents par pays dans le dataset **data**.
- Dataset "series" (EdStatsSeries.csv)
  - 3665 lignes, 21 variables, informations descriptives sur la façon dont sont produits les indicateurs du dataset **data** (description, périodicité, unité de mesure,…)

Description du jeu de données – 5 datasets (2/2)

- Dataset "data" (EdsStatscountry.csv)
  - 886930 lignes, 70 variables correspondant à la valeur de 3665 indicateurs par pays/année
  - 48 variables détaillant les valeurs des indicateurs de 1970 à 2017
  - 17 colonnes détaillant les projections des valeurs des indicateurs tous les 5 ans de 2020 à 2100

- Données significatives dans les datasets :
  - Variable « Region » (dataset country) pour regrouper les pays par régions
  - Variable « Topic » (dataset series) pour regrouper les indicateurs par catégories

=> Le dataset "data" est le dataset principal pour l'analyse pré-exploratoire des données, les autres datasets apportent peu d'informations.

Simplification du jeu de données

- Ajout de la variable "Topic" dans le dataset **data** provenant du dataset **series**

- Ajout de la variable « Region » dans le dataset **data** provenant du dataset **country**

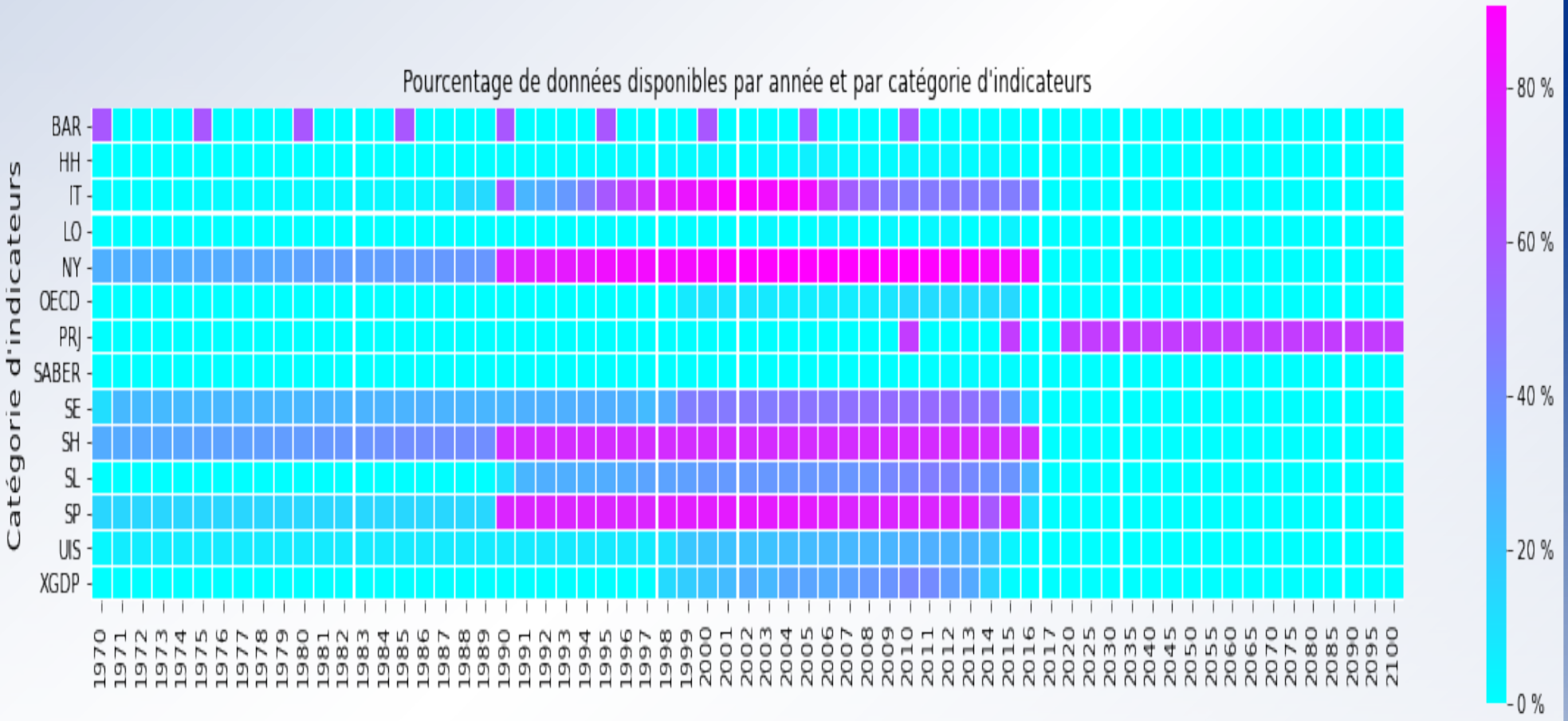
- Ajout de la colonne « Categ Indic » dans le dataset **data** pour regrouper les indicateurs selon les 3 premières lettres du code indicateur

=> les 4 autres dataset ne sont plus utiles (country, countryseries, series, footnote) dans le cadre de la problématique d'un projet d'expansion à l'international d'un Edtech

Sélection des indicateurs – 1er filtrage des données
Pourcentage des données disponibles par catégories d'indicateurs

- Recherche des données disponibles par catégories d'indicateurs/année(tous pays confondus) à partir du code indicateur
- But du 1er filtrage : éliminer des catégories d'indicateurs avec beaucoup de valeurs manquantes
- Pas de considération métier au niveau de ce filtre.
- Résultat : échantillon réduit à 40 % de total des indicateurs, 8 catégories d'indicateurs retenus **(IT, NY, SE, SH, SL, UIS, SP et XGDP)**

Sélection des indicateurs – pourcentage de données disponibles par année et catégories d'indicateurs



La majorité des données disponibles concerne la plage d'années entre 1990 et 2015.

Il y a peu de données disponibles en 2016 et presque pas en 2017.

Certaines catégories sont écartées car trop peu de données dispos (HH, LO, SABER,...)

Sélection des indicateurs – 2ème filtrage des données
Mise à l'écart des catégories métier non significatives pour l'étude

- Sélection des indicateurs sur des critères métier / fonctionnels
- Retrait des domaines d'indicateurs (variable « Topic ») non significatifs pour l'étude (exemples : «Pre-Primary», «Primary»)
- **Résultat du 2ème filtrage : l'échantillon des indicateurs (1199) est réduit à 32% des indicateurs totaux.**

Sélection des indicateurs – 3ème filtrage des données
Sélection de 11 indicateurs métier intéressants pour l'étude (1/2)

Prise en compte d'indicateurs dans la périmètre de l'étude :

- **IT - indicateurs sur l'utilisation et la pénétration d'Internet et des ordinateurs dans la population :**
  - IT.NET.USER.P2 (Internet users (per 100 people))
  - IT.CMP.PCMP.P2 (Personal computers (per 100 people))

- **NY - indicateurs sur les comptes nationaux, PIB, revenus par habitant :**
  - NY.GDP.PCAP.CD (GDP per capita (current US$))

- **XGDP - indicateurs sur les dépenses publiques dans l'éducation :**
  - XGDP.23.FSGOV.FDINSTADM.FFD (Government expenditure in secondary institutions education as % of GDP (%))
  - XGDP.56.FSGOV.FDINSTADM.FFD (Government expenditure in tertiary institutions as % of GDP (%))

- **SE/UIS - indicateurs divers sur l'éducation :**
  - SE.TER.ENRR (Gross enrolment ratio, tertiary, both sexes (%))

Sélection des indicateurs – 3ème filtrage des données
Sélection de 11 indicateurs métier intéressants pour l'étude (2/2)

(Suite de la liste...)

- **SE/UIS - indicateurs divers sur l'éducation :**
  - SE.SEC.ENRR.UP (Gross enrolment ratio,upper secondary,both sexes (%))
  - SE.ADT.1524.LT.ZS (Youth literacy rate, population 15-24 years, both sexes (%))
  - UIS.NER.3 (Net enrolment rate, upper secondary, both sexes (%))

- **SP pour les données sur la population :**
  - SP.TER.TOTL.IN (Population of the official age for tertiary education, both sexes (number))
  - SP.SEC.UTOT.IN (Population of the official age for upper secondary education, both sexes (number))

=> Les catégories « SH » et « SL » sont écartées car hors périmètre

=> Cette liste d'indicateurs va être filtrée par rapport au ratio de données disponibles par pays pour chaque indicateur

Sélection des indicateurs – 4ème filtrage des données
Sélection de pays avec au moins 50 % de données disponibles (1/2)

- Calcul taux moyen données manquantes par pays pour les 11 indicateurs sur la période 1990-2015

- Classement des pays par taux moyen NaN par ordre décroissant

- Récupération d'une liste de pays avec au moins 50 % de données disponibles

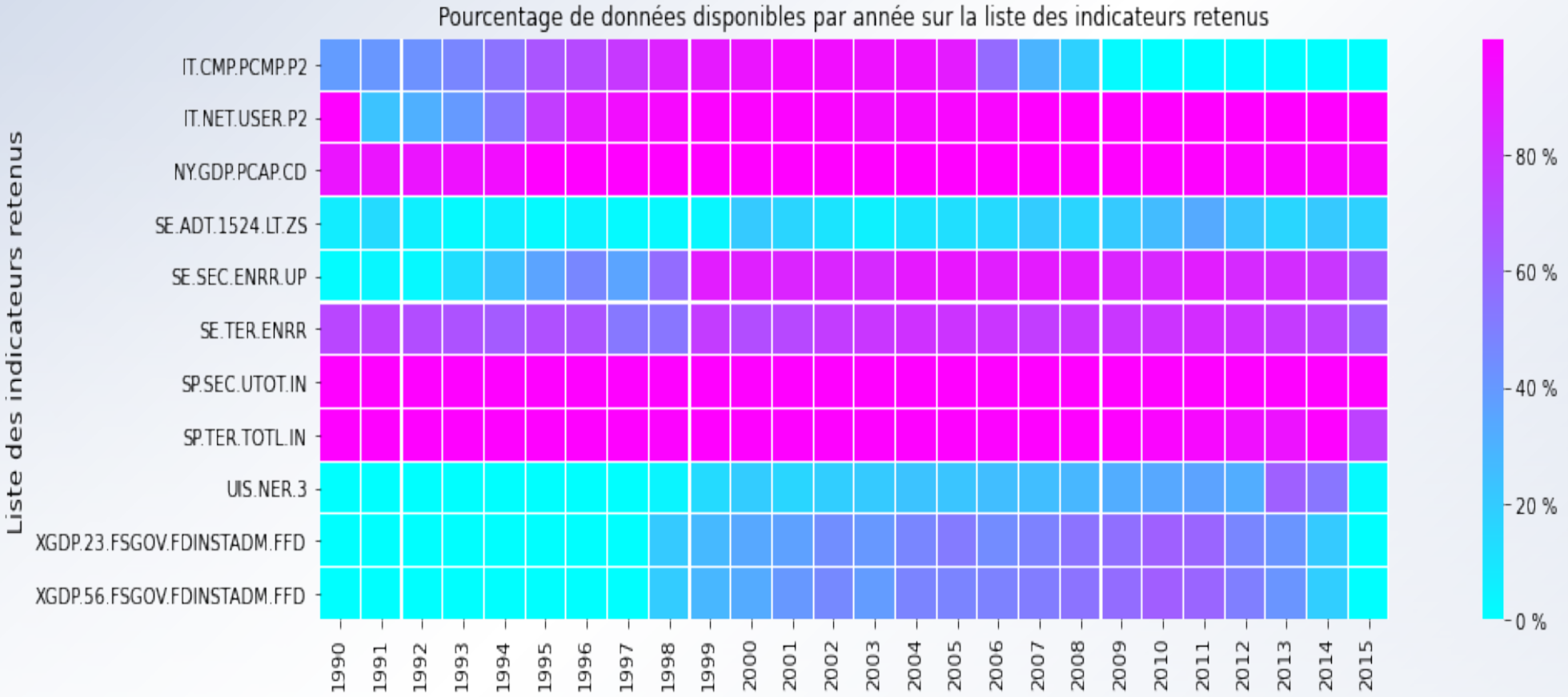
=> sélection de pays avec un seuil de données disponibles

- Calcul du pourcentage de données disponibles par année par indicateur sur la liste des pays retenus

=> écart de 3 indicateurs pour l'établissement du score d'attractivité

Sélection des indicateurs – 4ème filtrage de données

Sélection de pays avec au moins 50 % de données disponibles (2/2)



Les indicateurs SE.ADT.1524.LT.ZS et UIS.NER.3 ne sont pas retenus, trop peu de données disponibles sur la période 1990-2015.

L'indicateur IT.CMP.PCMP.P2 n’est pas retenu car trop peu de données disponibles de 2007 à 2015 (de 0 à 40 % de données disponibles).

Sélection des indicateurs – 5ème filtrage des données
Analyse de la distribution des données (1/2)

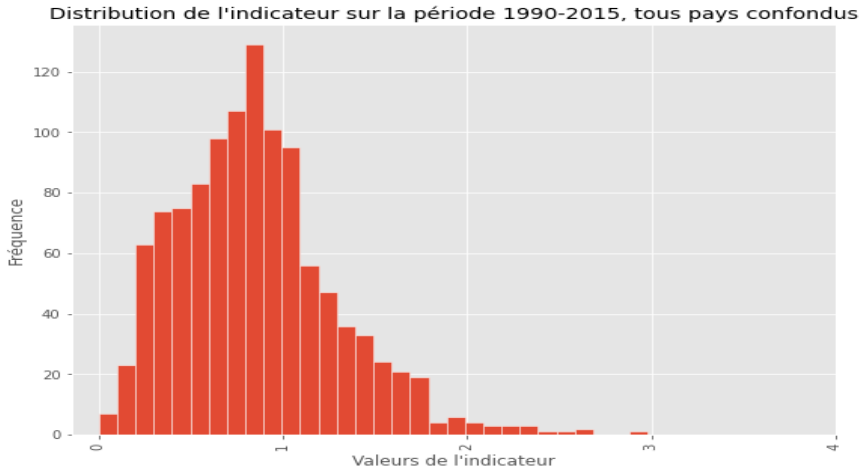
- Calcul moyenne, médiane et écart-type sur les 11 indicateurs

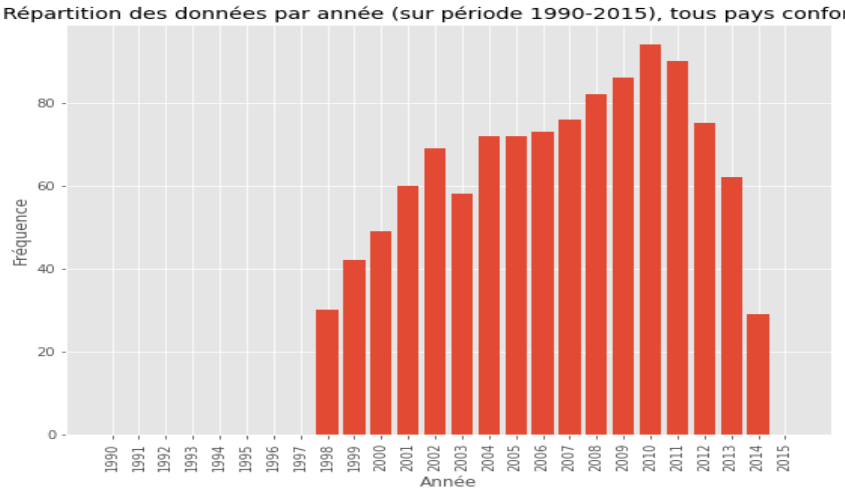
- Résultats principaux :
  - Distribution asymétrique (moyenne <> médiane) des données avec un échantillonnage continu des valeurs
  - Pas de « trous » dans les données sur la période 1990-2015
  - Outliers explicables, pas de valeurs aberrantes
  - Écarts-types importants indiquant une dispersion des données

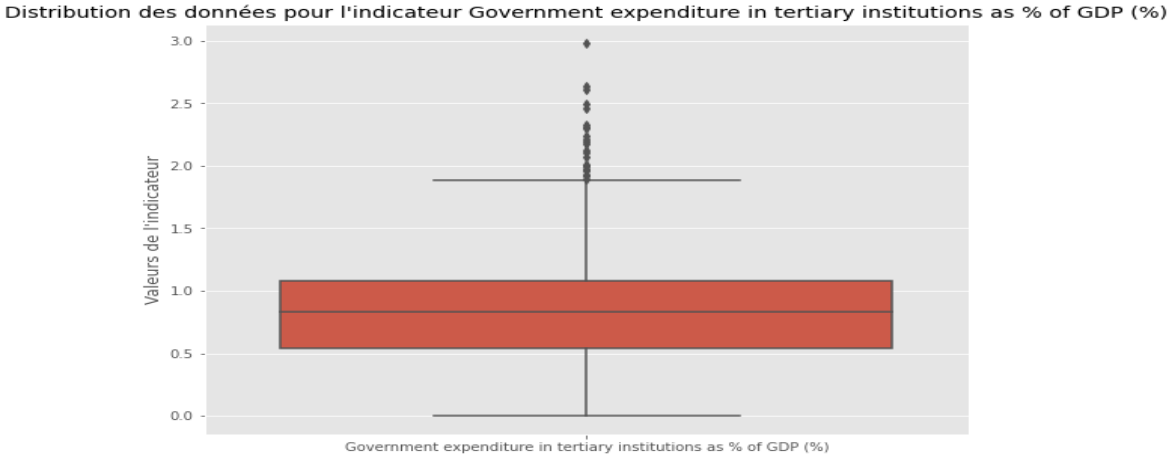
- A l'issue du 5ème filtrage, 6 indicateurs retenus
  - Les 2 indicateurs sur les dépenses sont écartés car manque de donnés récentes (trous sur l'année 2015)

Sélection des indicateurs – 5ème filtrage de données

Analyse de la distribution des données – Indicateurs de dépenses (2/2)







Moyenne sensiblement égale à la médiane, écart-type faible (peu de dispersion données)

Distribution des données faiblement asymétrique et « trous » dans la répartition des données sur la période avant 1998 et après 2014

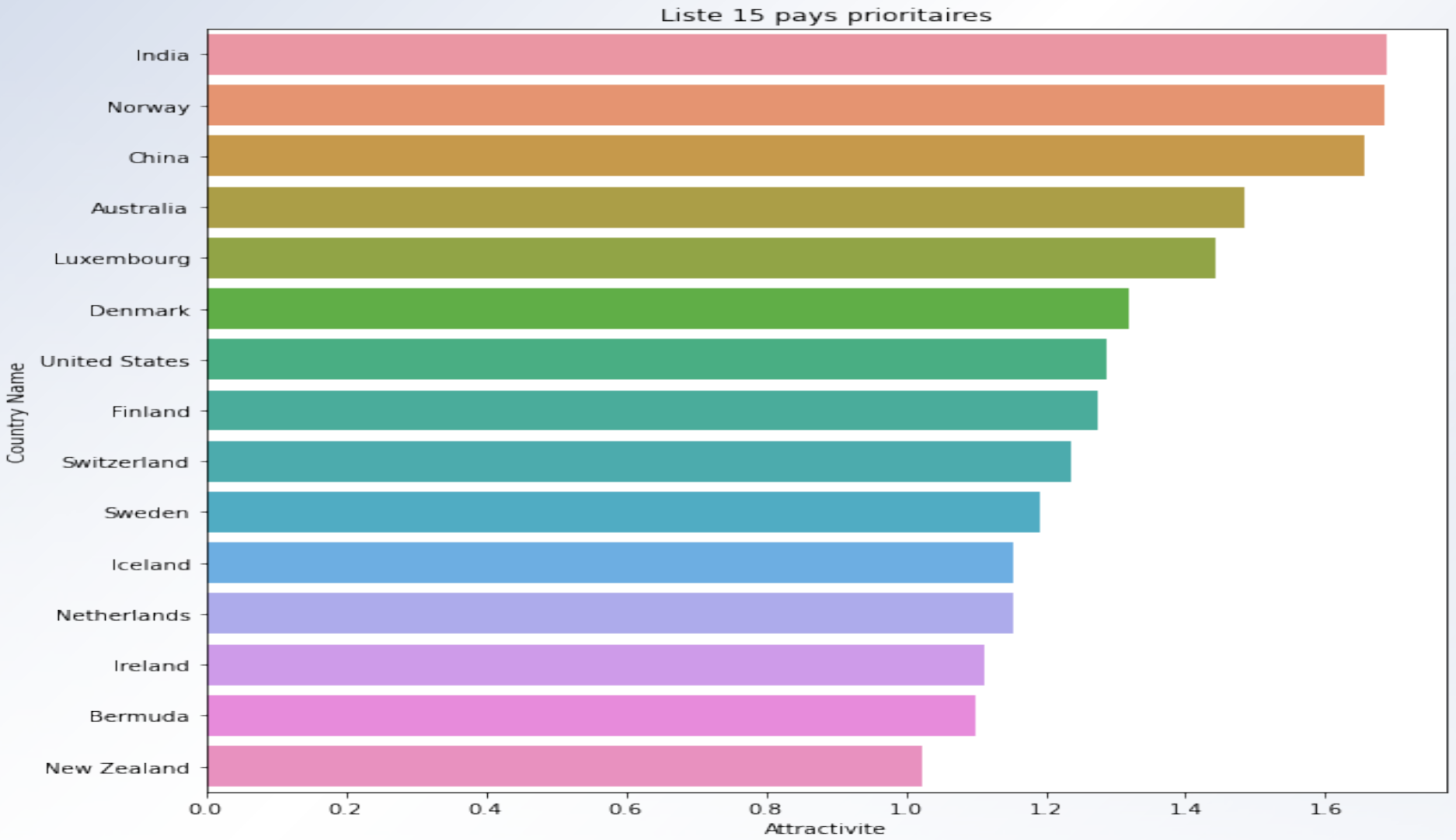
Démarches pour établir le modèle de scoring (1/2)

- Imputation des valeurs manquantes par la médiane, méthode robuste par rapport aux outliers
- Standardisation des données avec mise à l'échelle des valeurs car il est difficile de comparer des valeurs de population (nombre) avec des pourcentages
- Utilisation de la technique StandardScaler

Démarches pour établir le modèle de scoring (2/2)

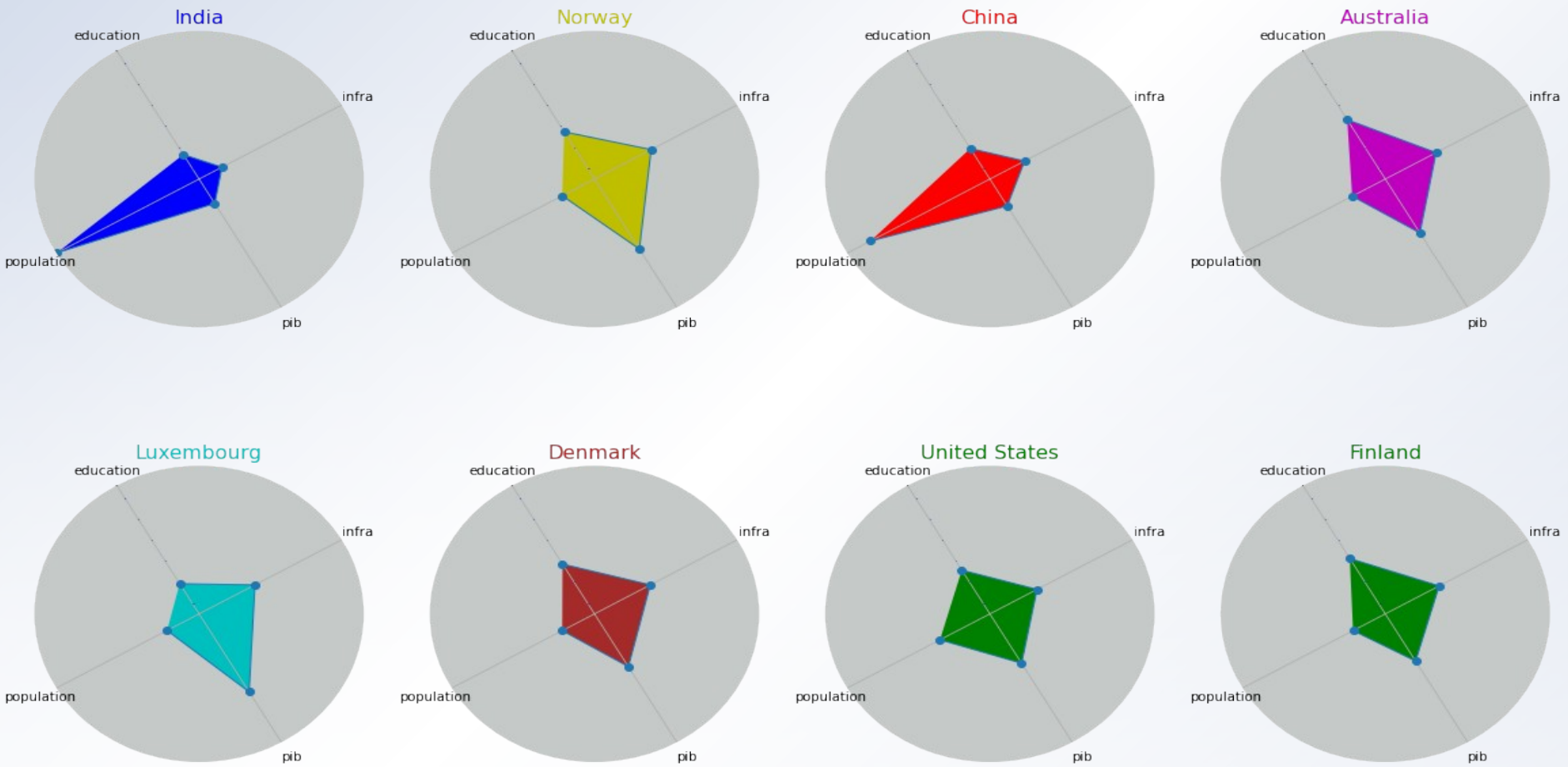
- Utilisation de 6 indicateurs classés par domaine : infrastructure internet (D1), revenus par habitant (D2), éducation (D3), population secondaire/tertiaire (D4)
- Période 2010-2015 retenue pour calculer un score avec des données récentes
- Score par domaine Dx = moyenne arithmétique valeurs indicateurs Dx / nombre d'années période 2010-2015
- Score global = moyenne arithmétique score Dx / nombre d'indicateurs Dx

Liste des 15 pays les plus prioritaires (score global)



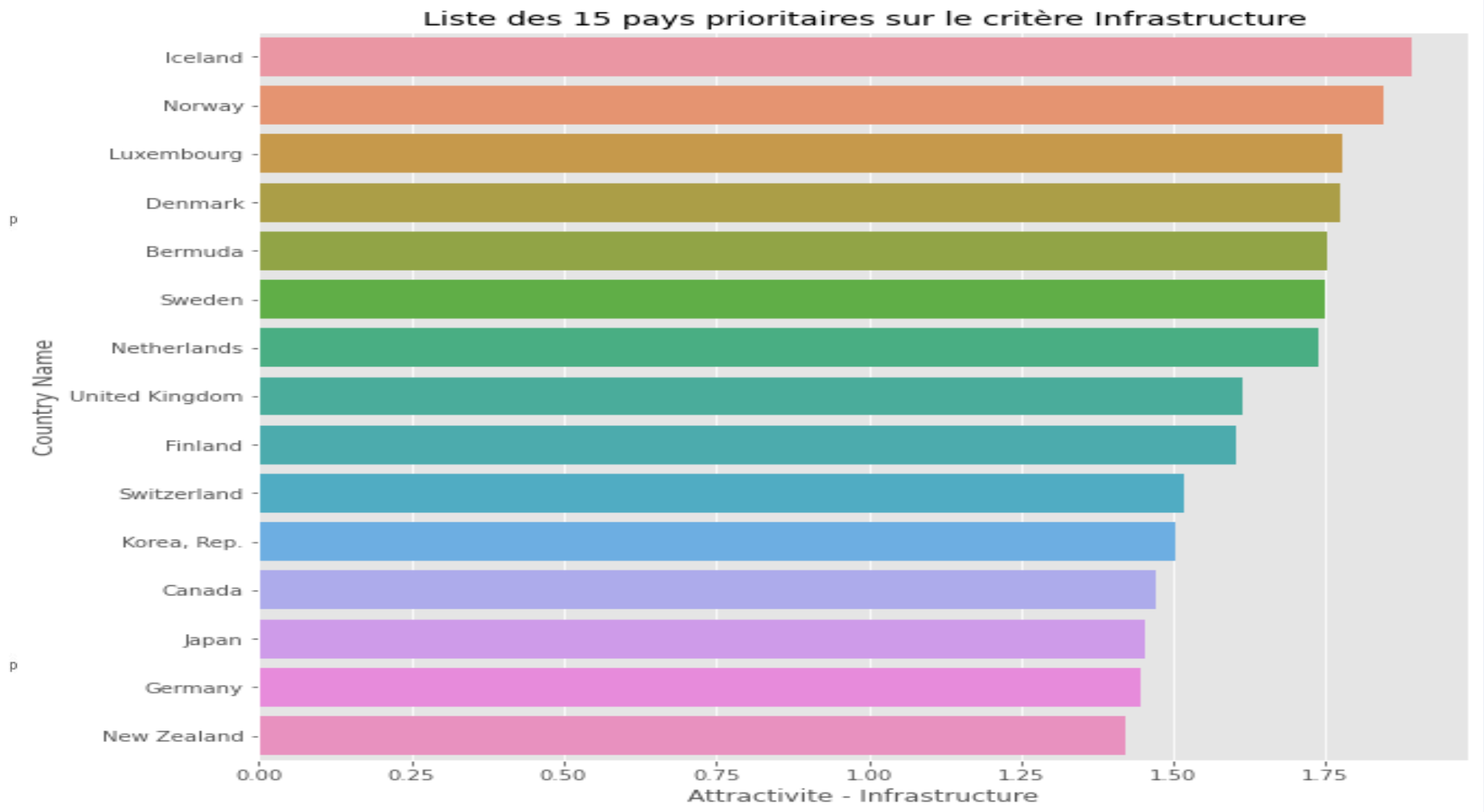
Pays les plus prioritaires dans le cadre de l'étude compte tenu des indicateurs :
Inde, Norvège, Chine, Australie, Luxembourg, Danemark, Etats-Unis, Finlande

Les 8 premiers pays pour le score global d'attractivité



Performances relativement homogènes sur les composantes du score global pour les Etats-Unis, l'Australie, le Danemark.
Des pays présentent des points forts: population pour la Chine ou l'Inde, revenu par habitant pour le Luxembourg ou la Norvège.

Score global vs Score par domaine des indicateurs



Le score global masque les forces/faiblesses des pays, nécessité d'étude des composantes du score global.
Ainsi, l'Islande possède un atout concernant la couverture Internet de sa population

Pistes d'amélioration du modèle de scoring

- Rendre le modèle de scoring paramétrable par l'utilisateur

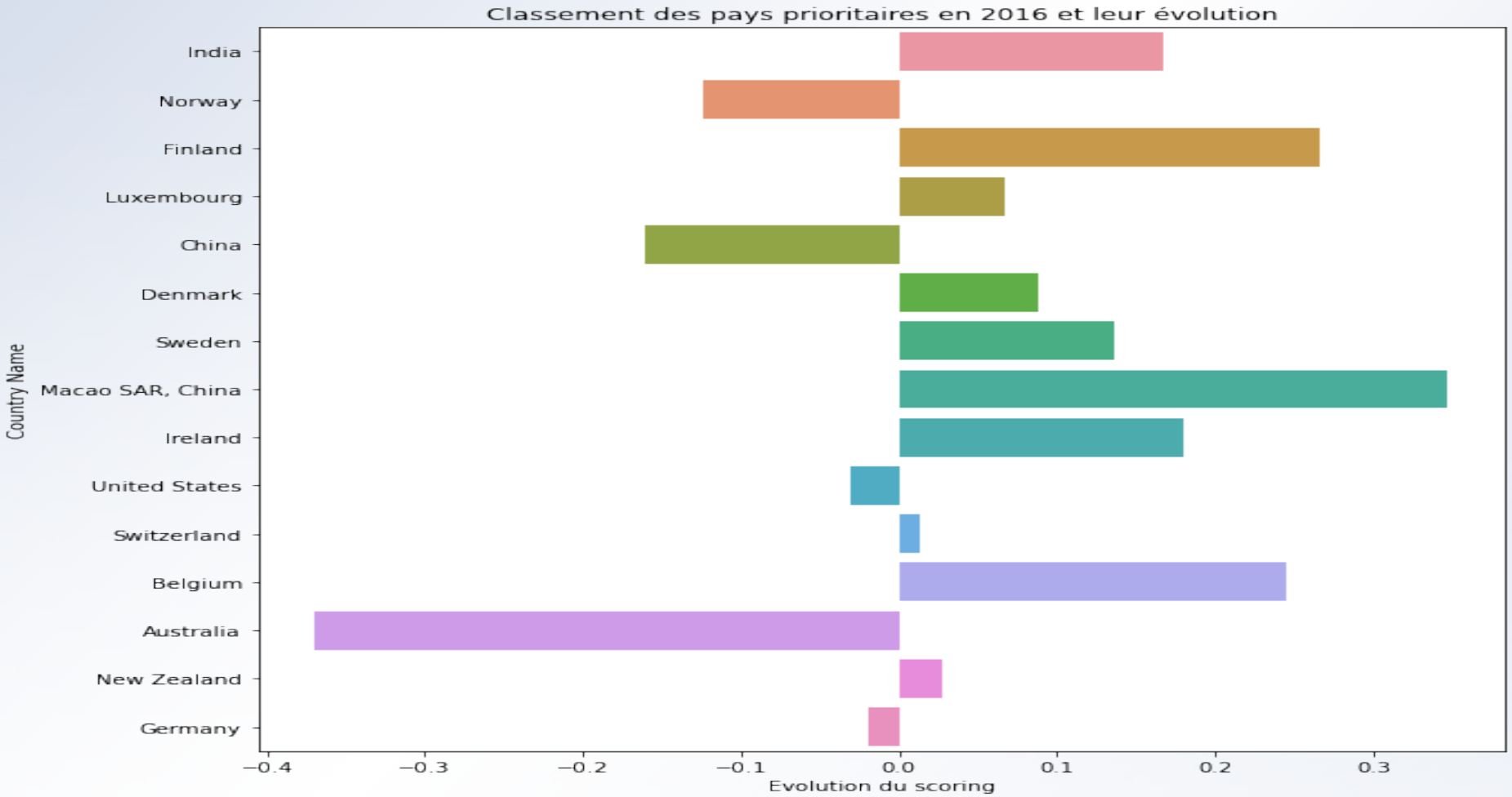
- Donner la possibilité de donner des poids différents à chaque domaine d'indicateurs permettant de calculer un score global à partir d'une moyenne pondérée des différents domaines/indicateurs

- Capacité à traiter facilement des nouvelles données en entrée du modèles de scoring.

Evolution du potentiel des pays : projection en 2016 (1/2)

- Calcul d'une régression linéaire à partir des scores globaux des pays sur la période 2010-2015 pour prédire l'évolution du score global en 2016.
- Calcul de l'évolution du potentiel entre la prédiction en 2016 et le score global sur la période 2010 à 2015.
- Utilisation de la méthode de régression linéaire, méthode simple à ce moment de la formation.
- Limites : vérifier la relation linéaire entre les composantes du scoring pour évaluer la pertinence de cette méthode.

Evolution du potentiel des pays : projection en 2016 (2/2)



Score +++ en 2016 : Macao, Finlande, Belgique, Inde

Score + (évol. modérée) en 2016 : Nouvelle-Zélande, Suisse, Luxembourg, Danemark

Score - - - en 2016 : Australie, Chine, Norvège

Score – (évol. modérée) en 2016 : Etats-Unis, Allemagne

Conclusion : pertinence du jeu de données - Atouts

- Prise en compte d'une liste exhaustive des pays
- Présence d'indicateurs significatifs mesurant les composantes de l'éducation au niveau lycées et universités:
  - indicateurs démographiques sur l'éducation
  - indicateurs sur les politiques d'investissement des pays
  - indicateurs sur des données macro-économiques (PIB par habitant, …)
- Les indicateurs sont tracés afin de donner la source des données

Conclusion : pertinence du jeu de données - Lacunes

- Le taux de valeurs manquantes très élevé sur la période 2010-2015 utilisée pour le scoring (environ 80 % sur la période)

- Absence d'éléments :
  - intérêt de la population pour la formation en ligne
  - dépenses sur Internet
  - proportion d'élèves se formant en dehors de leur établissement scolaire (soutien scolaire,...)

Annexes

Annexe 1 : liste des catégories d'indicateurs

Annexe 2 : liste des domaines des indicateurs pour scoring

Annexe 3 : Scoring - composante Education

Annexe 4 : Scoring - composante PIB

Annexe 5 : Scoring – composante Population

Annexe 6 : Etude relation entre indicateurs du scoring

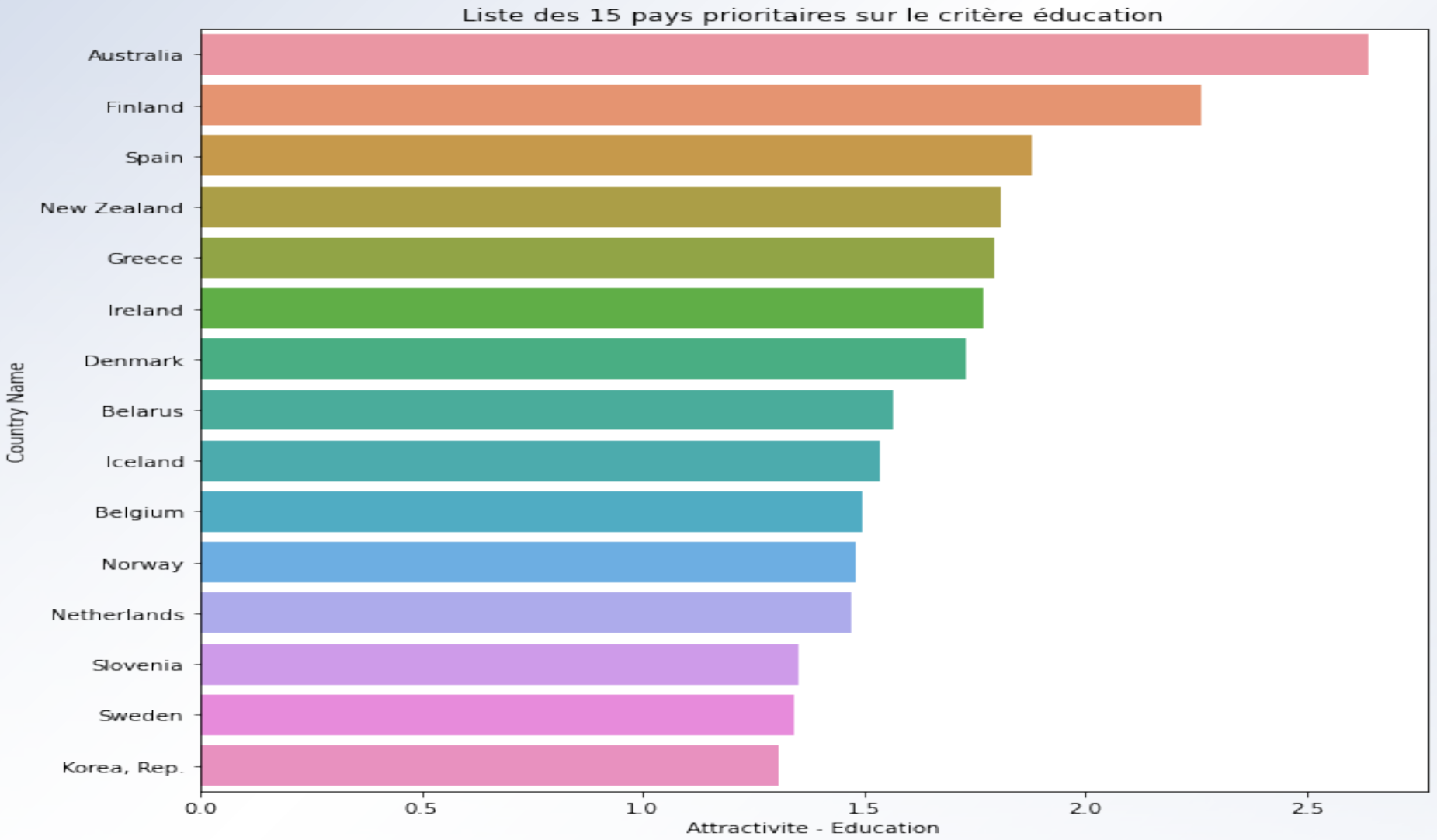
Annexe 1 : Liste des catégories d'indicateurs

- BAR : indicateurs Barro-Lee, distribution du niveau de scolarité, agrégés sur 5 ans selon 7 niveaux d'éducation, sur 144 pays à partir de 1970
- HH : indicateurs sur les enquêtes démographiques et sanitaires et sur des indicateurs multiples groupés
- IT : indicateurs sur l'utilisation et la pénétration d'Internet et des ordinateurs dans la population
- LO : indicateurs sur l'évaluation des apprentissages en lecture, science, mathématiques,…
- NY : indicateurs sur les comptes nationaux, PIB, revenus par habitant. Données macro-économiques.
- OECD : indicateurs sur les salaires des enseignants du secteur public dans l'OCDE
- PRJ : indicateurs à partir des projections Wittgenstein (durée de scolarisation, populations...)
- SABER : indicateurs mesurant les impacts sur la réforme des systèmes éducatifs au niveau des pays
- SE : indicateurs divers sur l'éducation
- SL : indicateurs sociaux sur la population active, sur le taux de chômage
- SH (Social Health) : indicateurs sur la santé au travail (taux de mortalité, risques au travail, …)
- SP : indicateurs sur la population (population active,…)
- UIS : indicateurs sur l'éducation de l'Unesco
- XGDP : indicateurs sur les dépenses publiques dans l'éducation

Annexe 2 : Domaines indicateurs retenus pour le scoring

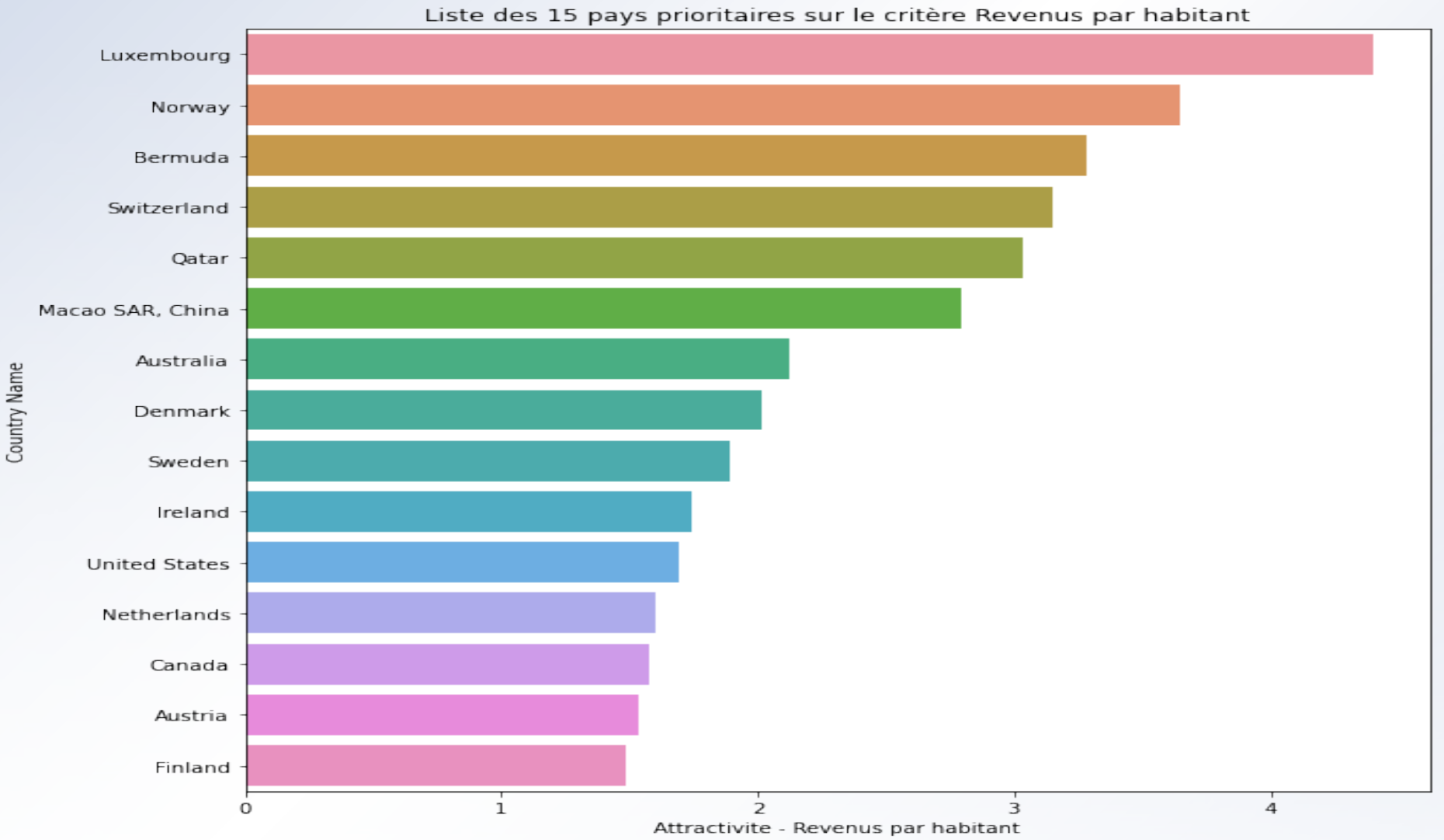
- **Domaine Education** : 2 indicateurs (taux brut de scolarisation dans l'enseignement tertiaire (universités) et pour le secondaire supérieur (lycées) en %
- **Domaine PIB** : 1 indicateur (PIB par habitant).
- **Domaine Infra** : 1 indicateur sur le nombre d'utilisateurs Internet (pour 100 personnes)
- **Domaine Population** : 2 indicateurs sur le nombre de personnes en âge pour entrer dans l'enseignement secondaire supérieur et dans l'enseignement supérieur à l'université.

Annexe 3 : Score - domaine Education



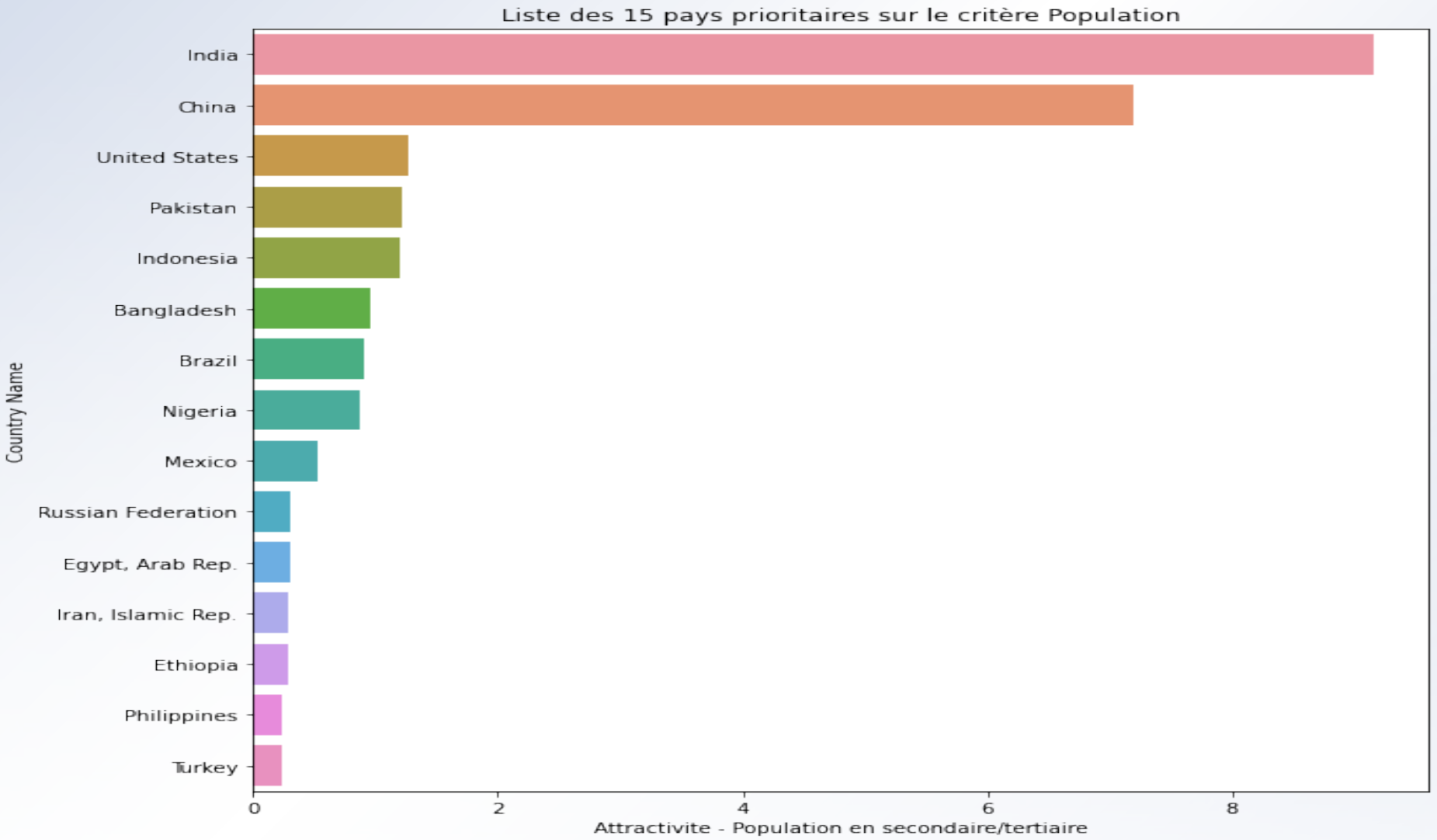
Dans le domaine Education, l'Australie, la Finlande, l'Espagne et la Nouvelle-Zélande ont les meilleurs taux brut de scolarisation dans le secondaire / enseignement supérieur.

Annexe 4 : Score - domaine PIB



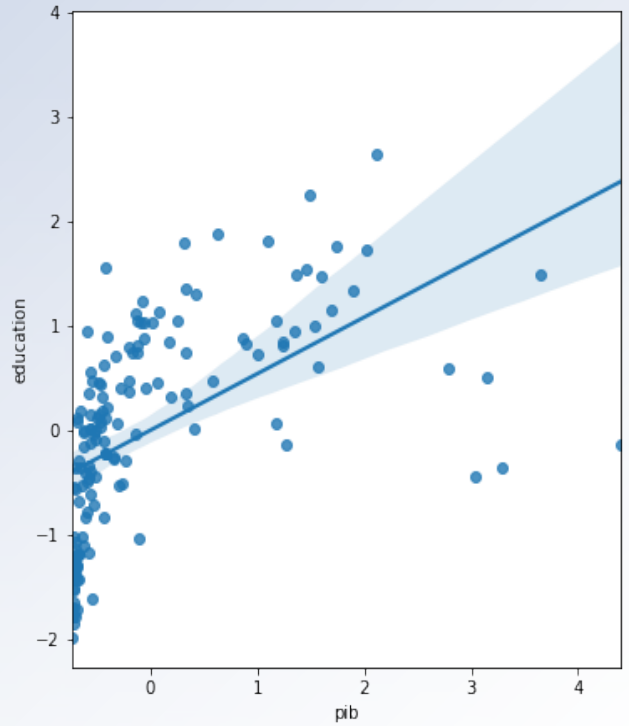
Dans le domaine PIB, les pays « riches comme le Luxembourg, la Norvège, la Suisse, le Qatar, ont le plus fort revenu par habitant.

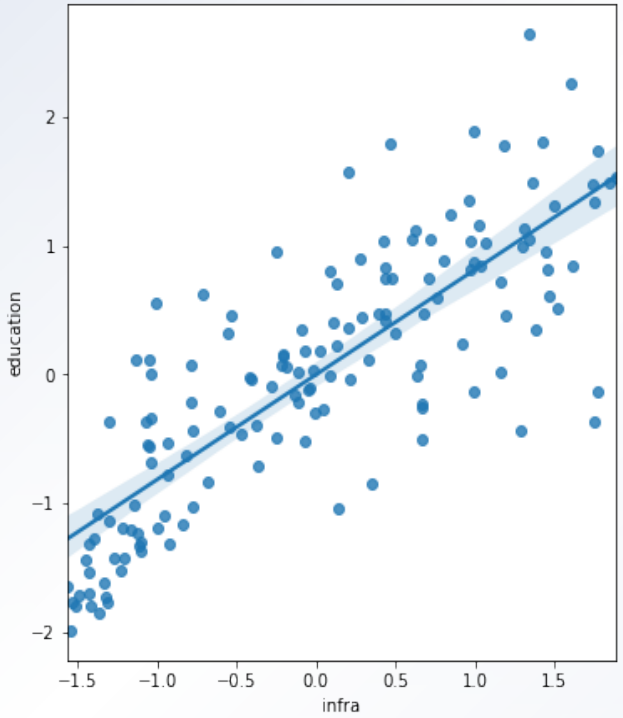
Annexe 5 : Score - domaine Population

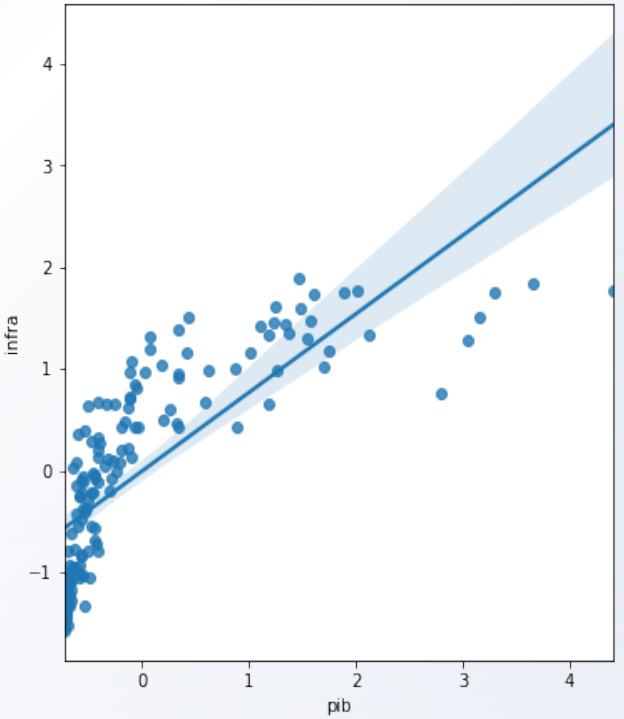


Dans le domaine Population, l'Inde, la Chine, et les Etats-Unis, sont les pays concentrant le plus de population en âge de suivre un enseignement en lycées ou universités.

Annexe 6 - Etude relations entre indicateurs du scoring







Pas de relation linéaire claire entre les domaines d'indicateurs constituant le score

Graphe du milieu : linéarité possible entre le taux de couverture internet et les pourcentages de scolarisation dans les lycées / universités.

Pertinence limitée de la méthode de régression linéaire si le lien linéaire entre les indicateurs n’est pas clairement établie.