Pré-analyse des données de la Banque Mondiale

academy

Problématique

Contexte : Projet d'expansion à l'international

Questions :

- Quels sont les pays avec un fort potentiel client?
- Pour chacun de ces pays, quel sera l'évolution de ce potentiel client?
- Dans quel pays l'entreprise doit-elle opérer en priorité?

Données à disposition :

- Utilisation des données de la Banque Mondiale pour 241 pays et régions du monde
- Plus de 3000 indicateurs décrivant l'accès à l'éducation, le PIB, la population...

Description du jeu de données

Les données de la Banque Mondiale contiennent 5 fichiers .csv :

- EdStatsCountry-Series
- EdStatsCountry
- EdStatsFootNote
- EdStatsSeries
- EdStatsData

Description du jeu de données (1)

- Le fichier **EdStatsCountry-Series** contient des informations sur le recueil des données pour différents indicateurs et pays.

Entrée [2]:

```
1 country_series = pd.read_csv("Edstats_csv/EdStatsCountry-Series.csv")
2 country_series.head()
```

Out[2]:

	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

Description du jeu de données (2)

- Le fichier **EdStatsFootNote** contient une description sur la source des données ou leurs incertitudes pour les différents pays en fonction de l'année.

Entrée [7]:

```
1 foot_note = pd.read_csv("Edstats_csv/EdStatsFootNote.csv")
2 foot_note.head()
```

Out[7]:

	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

Description du jeu de données (3)

- Le fichier **EdStatsCountry** contient des informations sur les différents pays et régions du monde répertoriés. La seule variable qui semble intéressante est « Income Group » qui classe les pays en différentes catégories selon leur richesse.

Entrée [13]:

```
1 countries = pd.read_csv("Edstats_csv/EdStatsCountry.csv")
2 countries.head()
```

Out[13]:

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	IMF data dissemination standard	p
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD	AW	...	NaN	
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income	AF	...	General Data Dissemination System (GDDS)	

Description du jeu de données (4)

- Le fichier **EdStatsSeries** contient des informations sur les 3665 indicateurs. Il y a plusieurs colonnes vides. Les colonnes les plus importantes sont le nom (Indicator Name) et les définitions (Short definition/Long definition). C'est dan ce fichier que nous allons chercher les indicateurs qui nous intéressent.

Entrée [19]:

```
1 indicators = pd.read_csv("Edstats_csv/EdStatsSeries.csv")
2 indicators.head(2)
```

Out[19]:

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 15-19 with...	Percentage of female population age 15-19 with...	NaN	NaN	NaN	NaN	NaN	...
1	BAR.NOED.1519.ZS	Attainment	Barro-Lee: Percentage of population age 15-19 ...	Percentage of population age 15-19 with no edu...	Percentage of population age 15-19 with no edu...	NaN	NaN	NaN	NaN	NaN	...

2 rows × 21 columns

Description du jeu de données (4)

- Le fichier **EdStatsSeries** contient des informations sur les 3665 indicateurs. Il y a plusieurs colonnes vides. Les plus importantes sont le nom (Indicator Name) et les définitions (Short definition/Long definition). C'est dan ce fichier que nous allons chercher les indicateurs qui nous intéressent.

Choix des variables intéressantes :

- L'accès à internet
- Le revenu par habitant
- La population cible
- Les prévisions futures de cette population.

1ère variable : l'accès à internet

2 indicateurs possibles : % de la population ayant un ordinateur personnel et % de la population utilisant internet. Ce deuxième indicateur contient des valeurs plus récentes et pour plus de pays que le premier, j'ai donc choisi de retenir celui-ci.

```
Entrée [93]:  1 indicators[indicators["Topic"] == "Infrastructure: Communications"]
```

Out[93]:

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...
610	IT.CMP.PCMP.P2	Infrastructure: Communications	Personal computers (per 100 people)	NaN	Personal computers are self-contained computer...	NaN	Annual	NaN	NaN	Weighted average	...
611	IT.NET.USER.P2	Infrastructure: Communications	Internet users (per 100 people)	NaN	Internet users are individuals who have used t...	NaN	Annual	NaN	NaN	Weighted average	...

2 rows × 21 columns

2ème variable : le revenu par habitant

Il y a plusieurs indicateurs qui représentent le PIB par habitant (GDP per capita) ou le RNB par habitant (GNI per capita). La différence réside dans la méthode de calcul.
J'ai choisi la méthode « Atlas » pratiquée par la Banque Mondiale elle-même qui exprime le revenu par habitant en dollar américain.

Entrée [94]:
```
1 indicators[indicators["Topic"] == "Economic Policy & Debt: National accounts: Atlas GN
```

Out[94]:

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...	Not fro origin sour
1668	NY.GNP.PCAP.CD	Economic Policy & Debt: National accounts: Atl...	GNI per capita, Atlas method (current US$)	NaN	GNI per capita (formerly GNP per capita) is th...	NaN	Annual	NaN	NaN	Weighted average	...	Na

3ème variable : la population

Il y a un indicateur qui exprime directement population du pays âgée entre 15 et 24 ans. Cela semble correspondre à la tranche d'âge adaptée aux services que nous proposons.

Entrée [30]:

```
1 pop = indicators[indicators["Indicator Name"] == "Population, ages 15-24, total"]
2 pop
```

Out[30]:

	Series Code	Topic	Indicator Name	Short definition	Long definition
2506	SP.POP.1524.TO.UN	Population	Population, ages 15-24, total	Population, ages 15-24, total is the total pop...	Population, ages 15-24, total is the total pop...

4ème variable : les prévisions de la population

Pour cette variable, j'ai choisi plusieurs indicateurs (voir ci-dessous). Un indicateur donne la population prévisionnelle d'une tranche d'âge donnée en fonction de son niveau d'étude.
J'ai donc créé un nouvel indicateur en sommant la valeur des 4 indicateurs ci-dessous.

```
Out[15]: array(['Projection: Population age 15-19 in thousands by highest level of educational attain
         ment. Upper Secondary. Total',
                'Projection: Population age 15-19 in thousands by highest level of educational attain
         ment. Post Secondary. Total',
                'Projection: Population age 20-24 in thousands by highest level of educational attain
         ment. Upper Secondary. Total',
                'Projection: Population age 20-24 in thousands by highest level of educational attain
         ment. Post Secondary. Total'],
               dtype=object)
```

Description du jeu de données (5)

- Le fichier **EdStatsData** contient les données numériques. Il y a la valeur des indicateurs pour tous les pays/régions du monde en fonction de l'année. Il y a beaucoup de données manquantes (86% du tableau est vide). Certaines données prévisionnelles sont estimées jusqu'en 2100.

Entrée [40]:

```
1 data = pd.read_csv("Edstats_csv/EdStatsData.csv")
2 data.head(2)
```

Out[40]:

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN

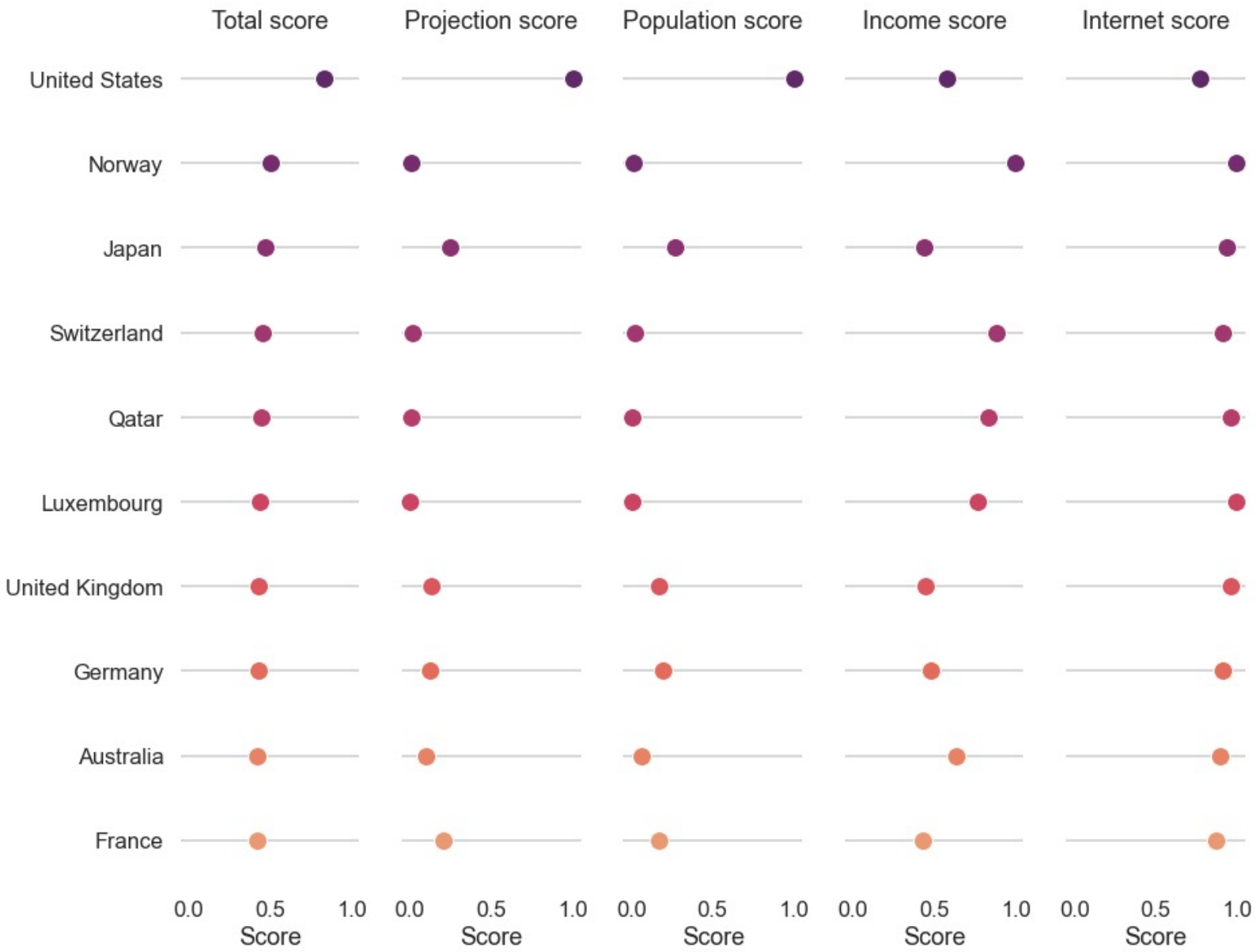
Nettoyage des données

- Filtrage des lignes :
  - Premier filtrage : sélection des pays riches uniquement. La sélection des pays riches réduit la liste des pays à 133.
  - Second filtrage : sélection des 7 indicateurs qui nous intéressent.
  - Troisième filtrage : suppression des pays ne n'ayant pas de valeur pour au moins un des 7 indicateurs. Il reste finalement **49** pays.

- Filtrage des colonnes:
  - Pour les variables internet, revenus et population on garde les données les plus récentes.
  - Pour la variable prévision de la population, on moyenne les valeurs jusqu'en 2040.

- Classement des pays :
  - Méthode de scoring : j'ai divisé les valeurs par la valeur max (score borné entre 0 et 1).
  - Score total : moyenne (non pondérée) des 4 scores.

Résultats



Les US arrivent premiers grâce à leur forte population. On peut voir qu'il y a un écart important entre les États-Unis et les autres pays sur les scores liés à la population.

Beaucoup de petits pays dans le haut du classement (Norvège, Luxembourg, Suisse, Qatar).

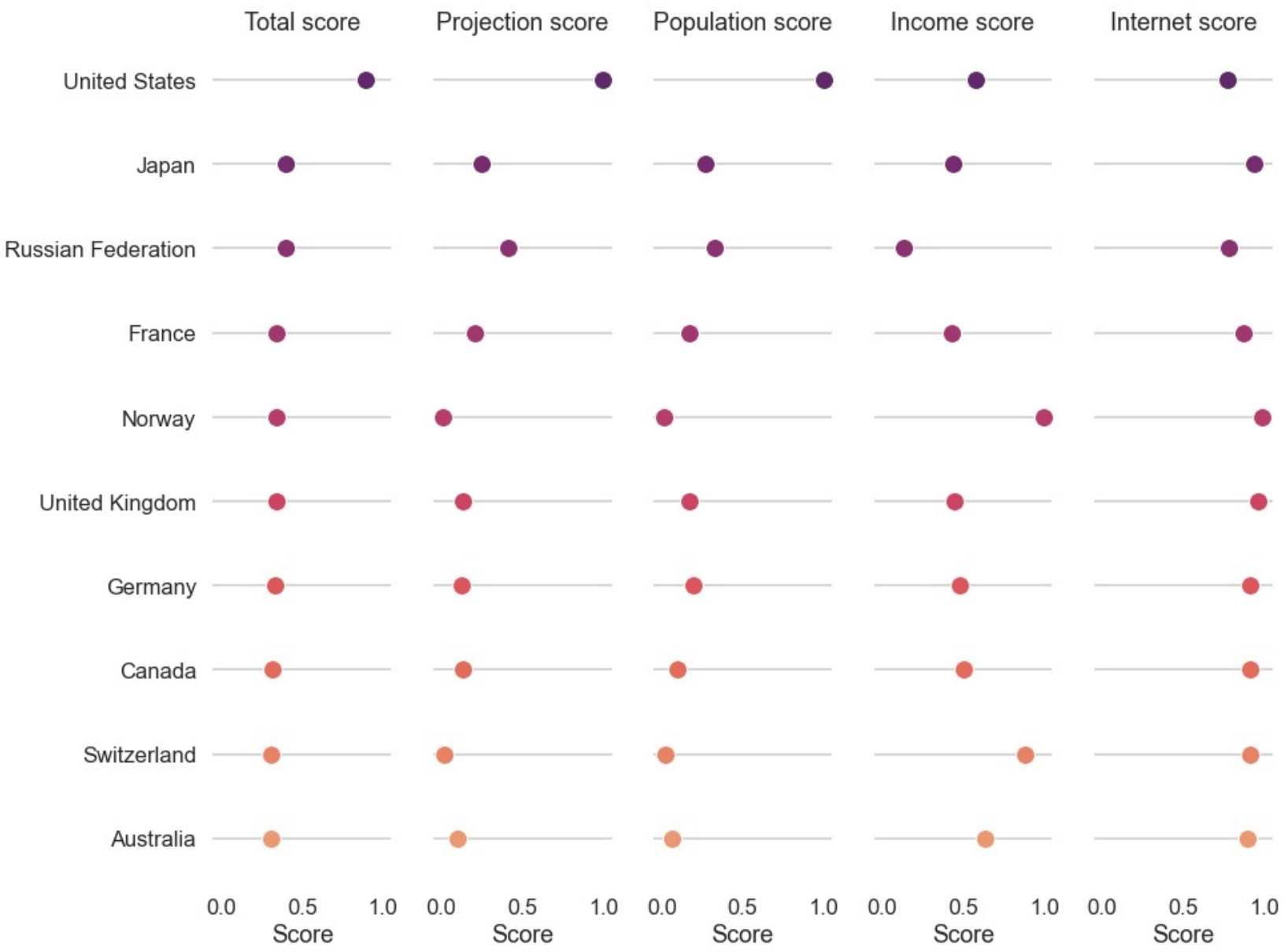
Autre méthode : moyenne pondérée

On peut envisager de calculer le score total en faisant une moyenne pondérée.

Les variables qui semblent les plus importantes ici sont Population et Projection. On va donc donner un poids plus important à ces variables.

J'ai arbitrairement choisi de donner un poids deux fois plus grands pour ces deux variables pour voir comment évolue le classement.

Résultats



Le Luxembourg et le Qatar sont sortis du top 10 et la Russie et le Canada l'ont intégré.

Le seul petit pays du top 5 est la Norvège.

Ce classement semble plus pertinent que le précédent.

La Chine n'apparaît pas dans le classement car elle ne fait pas partie des pays riches.

Perspectives

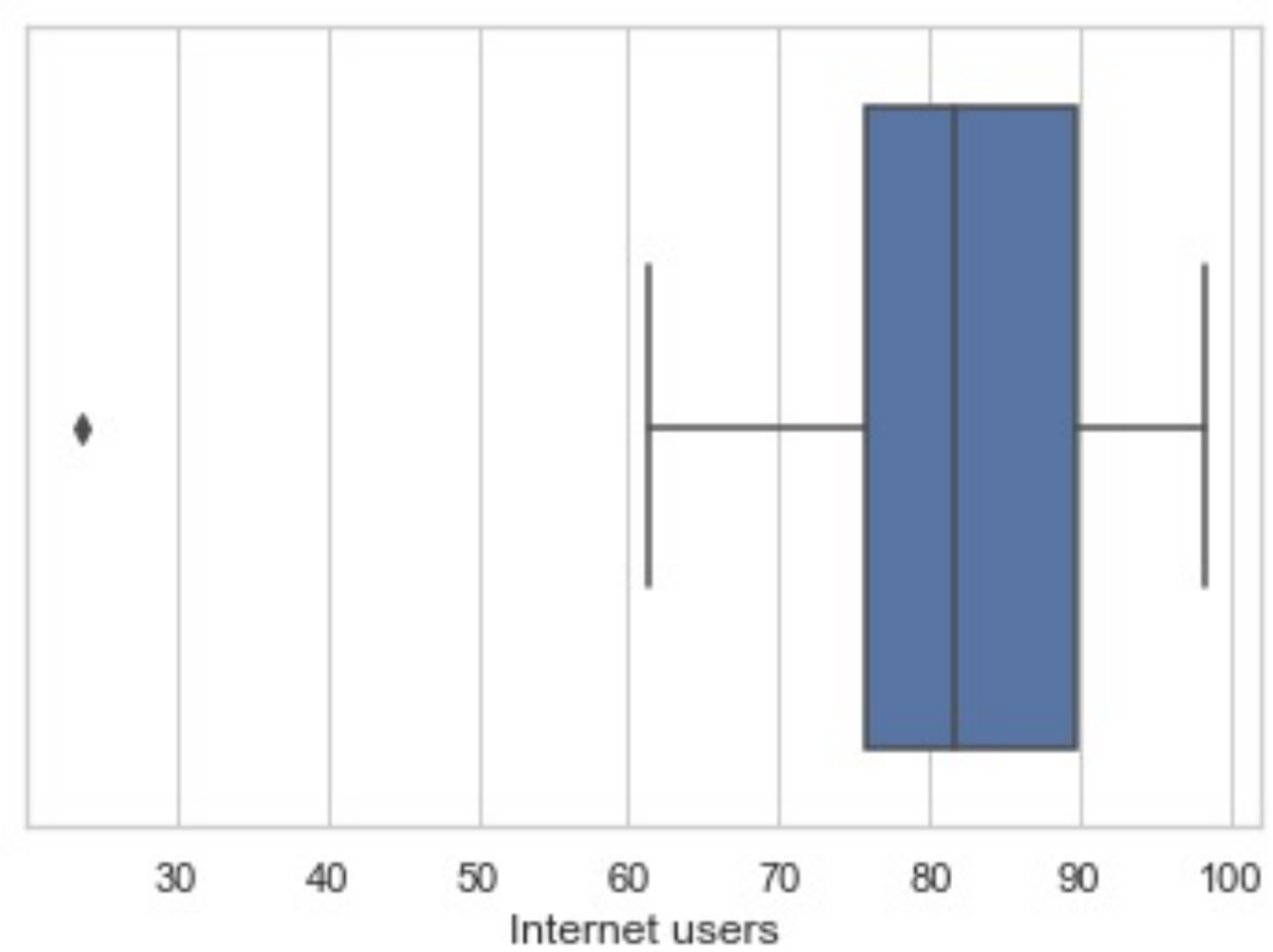
- Prendre en compte des pays qui ne sont pas « riches » mais qui peuvent quand même présenter un fort potentiel client, comme la Chine par exemple. Dans ce cas, il faudrait repenser à la méthode de scoring car la population de la Chine est tellement grande qu'elle rendrait le score des autres pays négligeable pour cet indicateur.

- On peut considérer d'autres indicateurs. Voir par exemple les indicateurs du sujet Engaging the Private Sector (SABER).

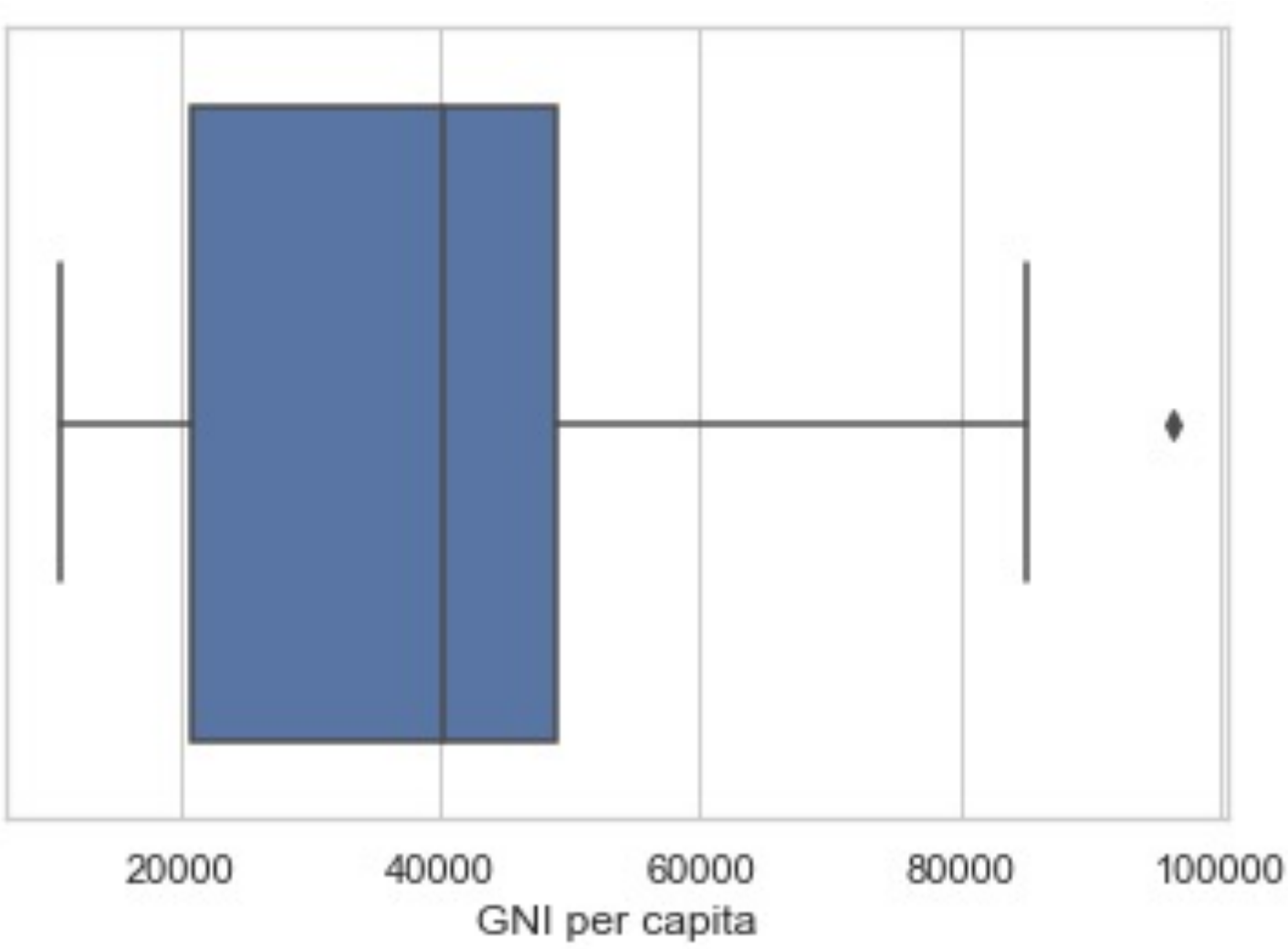
- Réfléchir aux poids à affecter aux différentes variables lors du calcul du score total.

Merci

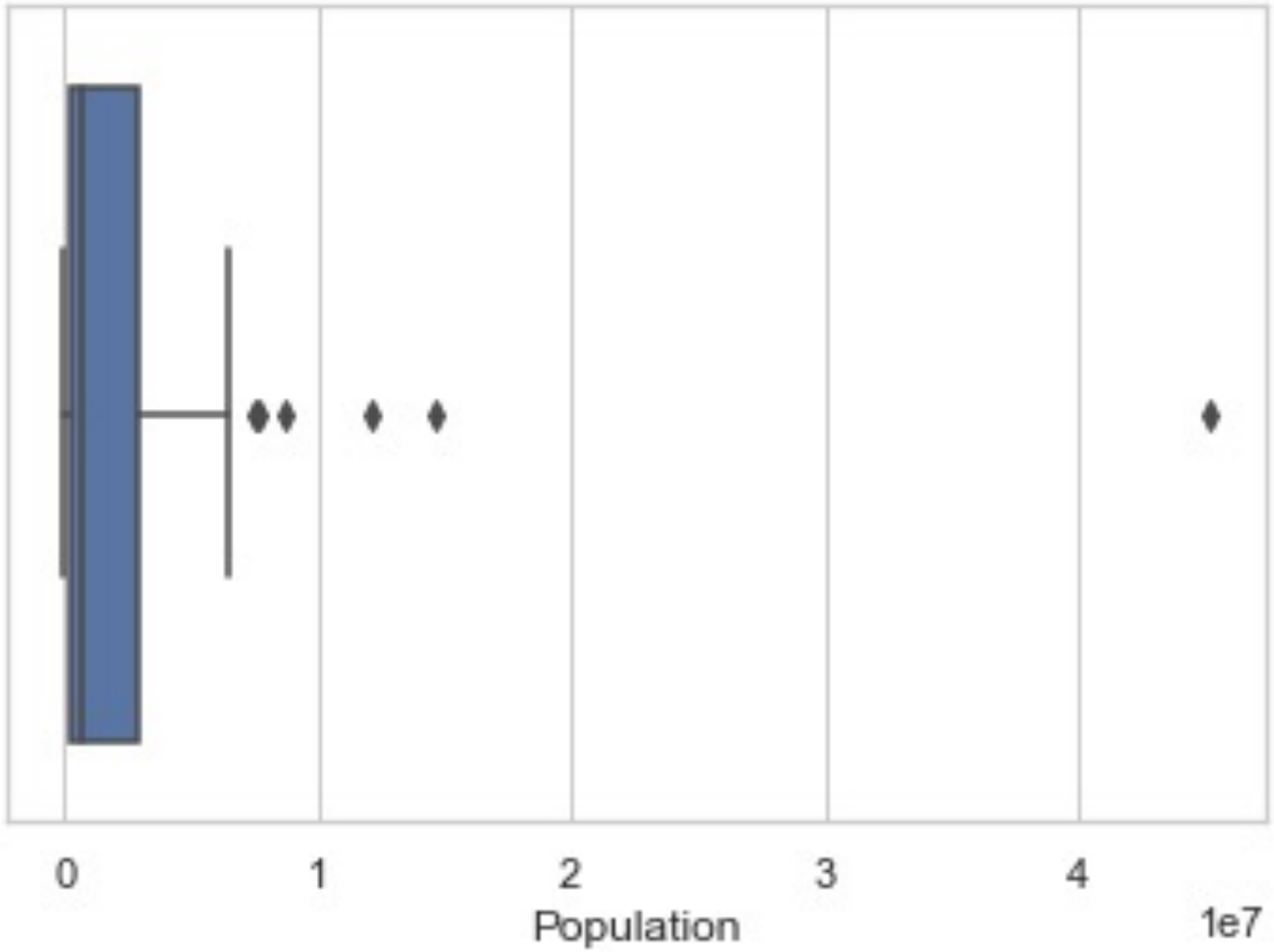
Distribution des valeurs Internet



Distribution des valeurs Income



Distribution des valeurs Population



Distribution des valeurs Projection