Analysez des données de systèmes éducatifs

Projet 2 – Parcours Data Scientist

Joachim CARON

Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

1 Objectifs

- Projet d'expansion à l'international de la start-up de la EdTech : *academy*

  -> **formation en ligne**

  -> **niveau lycée et université**

- Problématique :

  -> **Pays à fort potentiel de clients ?**

  -> **Evolution de ce potentiel de clients ?**

  -> **Pays dans lequel développer cette nouvelle offre en priorité ?**

- Cahier des charges :

  -> **Qualité** du jeu de données

  -> **Informations** contenues dans le jeu de données

  -> **Indicateurs pertinents**

  -> **Indicateurs statistiques classiques** pour les régions du monde et pays

1 Objectifs

- Outils à disposition

  -> Données sur l'éducation de la Banque mondiale : https://datacatalog.worldbank.org/dataset/education-statistics

  - **1 dataframe contenant les données numériques** : **« Data »**

  - 4 dataframe complémentaires : **« Country »**, **« Series »**, « Country-Series », « Footnote »

  -> Répertoire de **4000 indicateurs internationaux** concernant l'éducation : http://datatopics.worldbank.org/education/

Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

2 Présentation des jeux de données : Dataframe « Data »

- Description du jeu de données

-> **886 930 lignes :** 242 pays/régions/groupes × 3665 indicateurs

-> **70 colonnes** :

- 4 colonnes « object » : « Country Name », « Country Code », « Indicator Name », « Indicator Code »
- 65 colonnes « float » contenant les données numériques : années 1970 à 2100

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, both sexes (%)	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, female (%)	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2 Présentation des jeux de données : Dataframe « Data »

- Description du jeu de données

-> **Données manquantes**

- aucune dans les 4 colonnes « object »
- entre 72,6% et 99,9% de NaN dans les 65 colonnes contenant les données chiffrées
- 529525 lignes (soit 59,7% des combinaisons pays/indicateur) ne contenant que des NaN
    - Nombre non négligable de pays avec très peu de données (pays à priori non ciblés)
    - Peu de données pour les régions du monde
    - Nombre non négligeable d'indicateurs décrits mais inexploitables pour la suite de l'analyse

-> **Pas de duplicatas**

2 Présentation des jeux de données : Dataframe « Country »

- Description du dataframe « Country »

-> **241 lignes :** 241 pays/régions/groupes

- 1 pays de moins que dans le dataframe « data » -> British Virgin Islands (VGB)

-> **32 colonnes dont :**

- 4 colonnes « object » : « Country Code », « Short Name », « Table Name », « Long Name »
- 2 autres colonnes « object » : « Region » et « Income Group »
  - 7 régions
  - 5 groupes de revenus

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	National accounts base year	National accounts reference year	SNA price valuation	Lending category	Other groups
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from official government statistics; 1994-1999 from UN	Latin America & Caribbean	High income: nonOECD	AW	2000	NaN	Value added at basic prices	NaN	NaN

2 Présentation des jeux de données : Dataframe « Series »

- Description du dataframe « Series »

-> **3665 lignes :** 3665 indicateurs

- 53 indicateurs présents dans « Data » et absents de « Series » : exploitables si données
- 53 indicateurs présents dans « Series » et absents de « Data » : inexploitables
- 462 indicateurs avec Series Code / Indicator Code identique **mais** Indicator Name différent dans « Series » et « Data »

-> **21 colonnes dont :**

- colonnes « object » : « Series Code », « Topic », « Indicator Name », « Short definition », « Long definition »

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	Limitations and exceptions	Notes from original source	General comments
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age 15-19 with no education	Percentage of female population age 15-19 with no education	Percentage of female population age 15-19 with no education	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

3 Choix des indicateurs *via les descriptions sur http://datatopics.worldbank.org/education/*

Fort potentiel de clients

- Nombre d'étudiant potentiels élevé
- Rentabilité / Possibilité de payer
- Proportion d'étudiants dans le privé élevée

Formation **en ligne**

- Accès internet
- Tout le territoire national (pas de limitation géographique dans le pays)

Niveau **lycée et université**

- Etudiants dans le upper secondary et le tertiary (voire post-secondary non-tertiary)

3 Choix des indicateurs *via les descriptions sur http://datatopics.worldbank.org/education/*

Fort potentiel de clients

- Nombre d'étudiant potentiels élevé
- Rentabilité / Possibilité de payer → GNI per capita, PPP (current international $)
- Proportion d'étudiants dans le privé élevée →
  - Percentage of enrolment in tertiary education in private institutions (%)
  - Percentage of enrolment in post-secondary non-tertiary education in private institutions (%)
  - Percentage of enrolment in upper secondary education in private institutions (%)

Formation **en ligne**

- Accès internet → Internet users (per 100 people)
- Tout le territoire national (pas de limitation géographique dans le pays)

Niveau **lycée et université**

- Etudiants dans le upper secondary et le tertiary (voire post-secondary non-tertiary) →
  - Enrolment in upper secondary education, both sexes (number)
  - Enrolment in tertiary education, all programmes, both sexes (number)
  - Enrolment in post-secondary non-tertiary education, both sexes (number)

3 Choix des indicateurs *via les descriptions sur http://datatopics.worldbank.org/education/*

Evolution du potentiel de clients

- Projection de l'évolution du nombre d'étudiants diplômés jusqu'à 2100

3 Choix des indicateurs *via les descriptions sur http://datatopics.worldbank.org/education/*

Evolution du potentiel de clients

- Projection de l'évolution du nombre d'étudiants diplômés jusqu'à 2100

→ Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total

→ Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total

Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL : *Régions du Monde*
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

4 Analyse approfondie du jeu de données principal : *Régions du Monde*

1) Création d'un dataframe contenant :

-> les **7 régions du monde** : « Latin America & Caribbean », « South Asia », « Sub-Saharan Africa », « Europe & Central Asia », « Middle East & North Africa », « East Asia & Pacific », « North America »

-> les **indicateurs sélectionnés**

puis **suppression des colonnes et des lignes ne contenant que des NaN**

- Aucunes données pour indicateurs suivants : Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total

  Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total

4 Analyse approfondie du jeu de données principal : *Régions du Monde*

2) Création d'un dataframe **pour chaque indicateur** de la liste finale puis sélection des **données de la dernière année connue**

3) Dataframes fusionnés par jointures internes successives

4) Création d'une nouvelle colonne **« Enrolment upper secondary and tertiary »** résultant de la **somme des 3 indicateurs « Enrolment »**

5) Création d'une nouvelle colonne en **divisant les variables de la nouvelle colonne « Enrolment upper secondary and tertiary » par le nombre de pays par région**

	Region	GNI per capita (2016)	Internet users (2016)	Enrolment tertiary (2014)	Enrolment upper secondary (2014)	Enrolment post-secondary (2014)	Enrolment private post-secondary (2014)	Enrolment upper secondary and tertiary	Nombre de pays	Enrolment divided by number of countries
0	East Asia & Pacific	17023.396558	52.796321	69097800.0	68691424.0	3231800.0	41.461529	141021024.0	32	4406907.0
1	Europe & Central Asia	30824.645377	73.914751	37693412.0	35525304.0	2033528.5	40.008018	75252244.5	54	1393560.083333
2	Latin America & Caribbean	15026.767081	56.373113	24087544.0	25425662.0	237836.734375	38.706848	49751042.734375	37	1344622.776605
3	Middle East & North Africa	19619.017553	47.621417	14332233.0	15030861.0	1745216.75	3.36045	31108310.75	18	1728239.486111
4	North America	57163.053631	77.540410	21096660.0	13246597.0	1499981.5	37.9884	35843238.5	3	11947746.166667
5	South Asia	6062.797051	26.466113	33412020.0	69851160.0	900454.375	39.426479	104163634.375	8	13020454.296875
6	Sub-Saharan Africa	3612.759512	19.998498	7795920.5	21157512.0	1338692.625	17.28047	30292125.125	43	704468.026163

4 Analyse approfondie du jeu de données principal : *Régions du Monde*

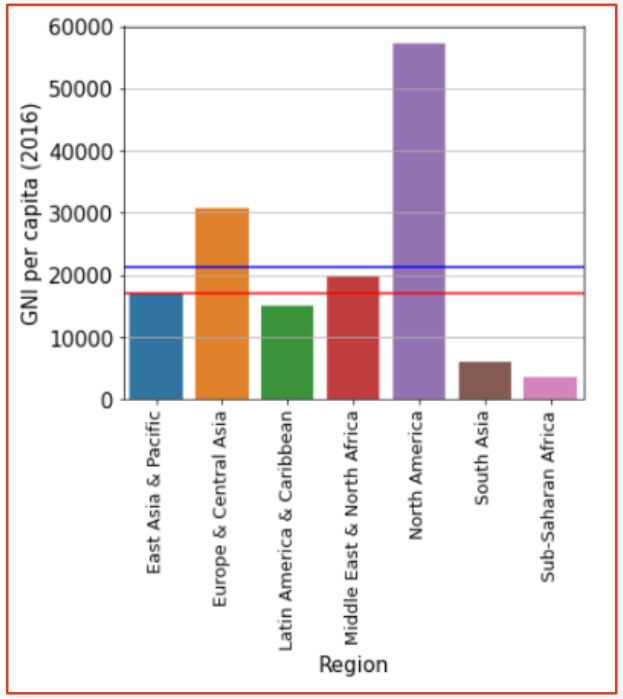
7) Grandeurs statistiques

GNI per capita

Moyenne ≈ 21 333 PPP$

Médiane ≈ 17 023 PPP$

Ecart-type ≈ 18 177 PPP$

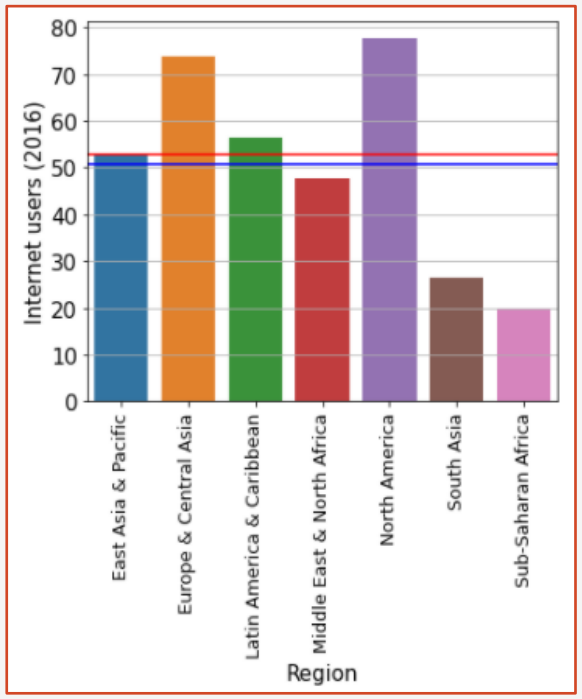


Internet Users

Moyenne ≈ 50,7%

Médiane ≈ 52,8%

Ecart-type ≈ 21,7%

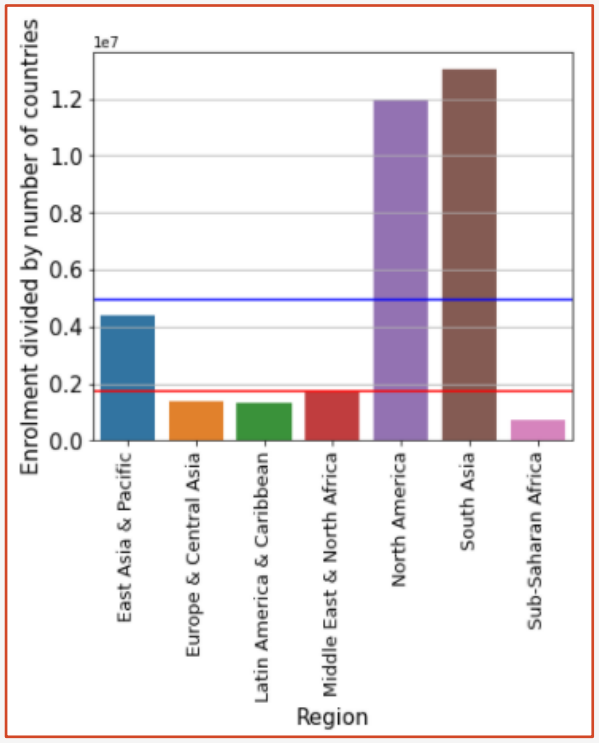


Enrolment divided by number of countries

Moyenne ≈ 4 935 143

Médiane ≈ 1 728 239

Ecart-type ≈ 5 298 512

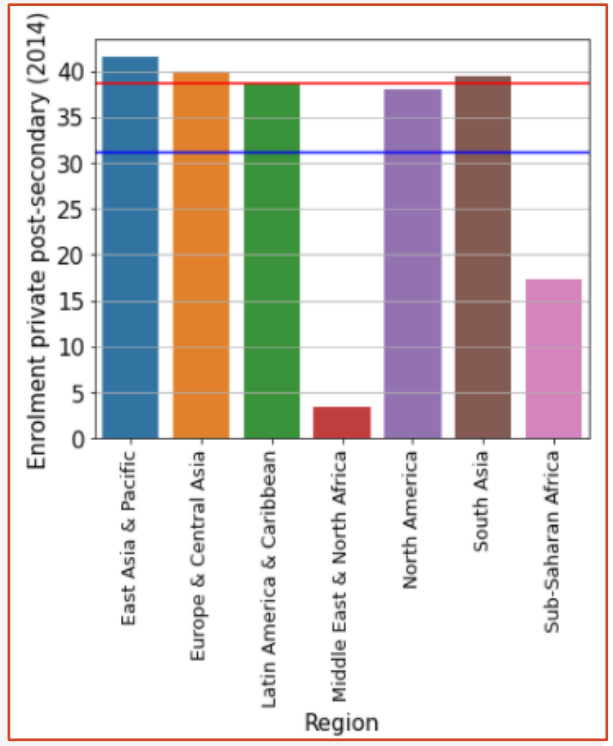


Enrolment private post-secondary

Moyenne ≈ 31.2%

Médiane ≈ 38,7%

Ecart-type ≈ 14,8%



4 Analyse approfondie du jeu de données principal : *Régions du Monde*

8) SCORING :

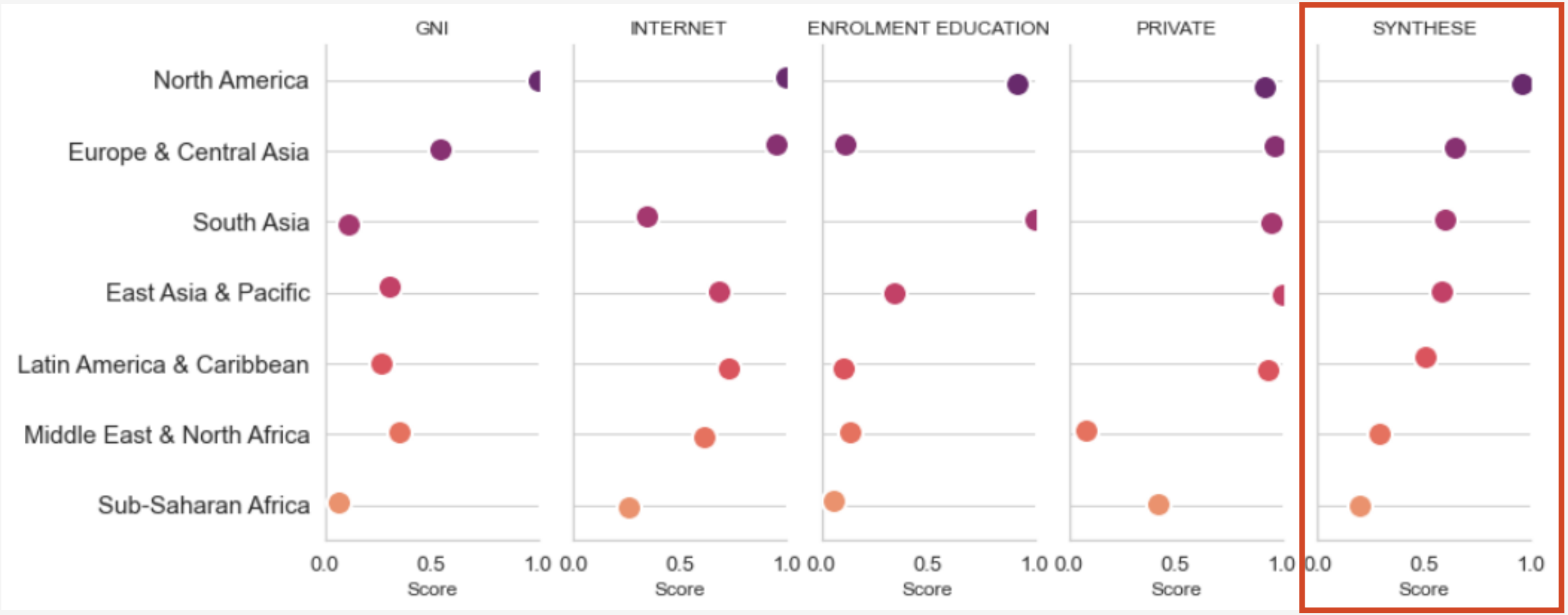
-> Le score de chaque variable de chaque indicateur est ramené entre 0 et 1 (chaque variable est divisée par la valeur maximale de chaque colonne)

-> Chaque indicateur pèse le même poids (25%) pour le score final

	Region	SCORE_GNI	SCORE_INTERNET	SCORE_ENROLMENT	SCORE_PRIVATE	SCORE_SYNT
4	North America	1.000000	1.000000	0.917614	0.916232	0.958462
1	Europe & Central Asia	0.539241	0.953242	0.107029	0.964943	0.641114
5	South Asia	0.106061	0.341320	1.000000	0.950917	0.599575
0	East Asia & Pacific	0.297804	0.680888	0.338460	1.000000	0.579288
2	Latin America & Caribbean	0.262876	0.727016	0.103270	0.933561	0.506681
3	Middle East & North Africa	0.343212	0.614150	0.132733	0.081050	0.292786
6	Sub-Saharan Africa	0.063201	0.257911	0.054105	0.416783	0.198000

4 Analyse approfondie du jeu de données principal : *Régions du Monde*

9) VISUALISATION DU SCORING :



Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL : *Pays*
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

4 Analyse approfondie du jeu de données principal : *Pays*

1) Création d'un dataframe contenant les **pays uniquement** et les **indicateurs sélectionnés**

Liste finale des indicateurs pour les pays :

- GNI per capita, PPP (current international $)
- Internet users (per 100 people)
- Enrolment in upper secondary and tertiary education
  - Enrolment in upper secondary education, both sexes (number)
  - Enrolment in tertiary education, all programmes, both sexes (number)
  - Enrolment in post-secondary non-tertiary education, both sexes (number)
- Private upper secondary and tertiary
  - Percentage of enrolment in post-secondary non-tertiary education in private institutions (%)
  - Percentage of enrolment in tertiary education in private institutions (%)
  - Percentage of enrolment in upper secondary education in private institutions (%)

Utilisés pour la partie évolution du potentiel de clients

- Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total
- Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total

4 Analyse approfondie du jeu de données principal : *Pays*

2) Création d'un dataframe pour chacun des indicateurs puis nettoyage des données : suppression des colonnes puis des lignes ne contenant que des NaN

-> On ne garde que les données de l'année la plus récente

-> Un **filtre à 1 millions d'étudiants** est appliqué à la nouvelle colonne « Enrolment upper secondary and tertiary »

- Il reste une sélection de 57 pays pour la suite de l'analyse

-> On ne garde pas les indicateurs « Enrolment in post-secondary non-tertiary education, both sexes (number) » et « Percentage of enrolment in post-secondary non-tertiary education in private institutions (%) » car 71 pays supprimés (sur 217)

3) Création d'une nouvelle colonne « Private upper secondary and tertiary » résultant de la moyenne entre « Private upper secondary » et « private tertiary

4 Analyse approfondie du jeu de données principal : *Pays*

4) Création du dataframe final en fusionnant tous les dataframe par jointures internes successives

	Country Name	Enrolment upper secondary	Enrolment tertiary	Enrolment upper secondary and tertiary	Private upper secondary	Private tertiary	Private upper secondary and tertiary	Internet users	GNI
0	Afghanistan	968769.00	262874.00	1231643.00	1.6171	41.783138	21.700119	10.595726	1900.0
1	Algeria	1332584.00	1289474.00	2622058.00	0.3142	0.0	0.1571	42.945527	14420.0
2	Argentina	1722700.00	2869450.00	4592150.00	29.621519	25.77264	27.69708	70.150764	19530.0
3	Australia	1104162.00	1453521.00	2557683.00	40.600109	8.54195	24.57103	88.238658	45970.0
4	Bangladesh	5784713.50	2068355.00	7853068.50	90.423538	43.026951	66.725245	18.246938	3790.0
5	Belgium	776413.00	504745.00	1281158.00	59.17572	56.641079	57.9084	86.5165	46010.0
6	Bolivia	665372.00	352554.00	1017926.00	14.69449	19.52779	17.11114	39.697515	7120.0
7	Brazil	9949583.00	8285475.00	18235058.00	13.7564	73.901863	43.829132	59.682747	14840.0
8	Canada	1531393.00	1212161.00	2743554.00	6.02197	0.0	3.010985	89.84	43420.0
9	Chile	1032041.00	1221774.00	2253815.00	63.031601	84.645889	73.838745	66.01	23290.0
10	China	43709224.00	43367392.00	87076616.00	10.05704	13.53814	11.79759	53.2	15500.0
11	Colombia	1334617.00	2293550.00	3628167.00	24.28524	49.07946	36.68235	58.136494	13920.0

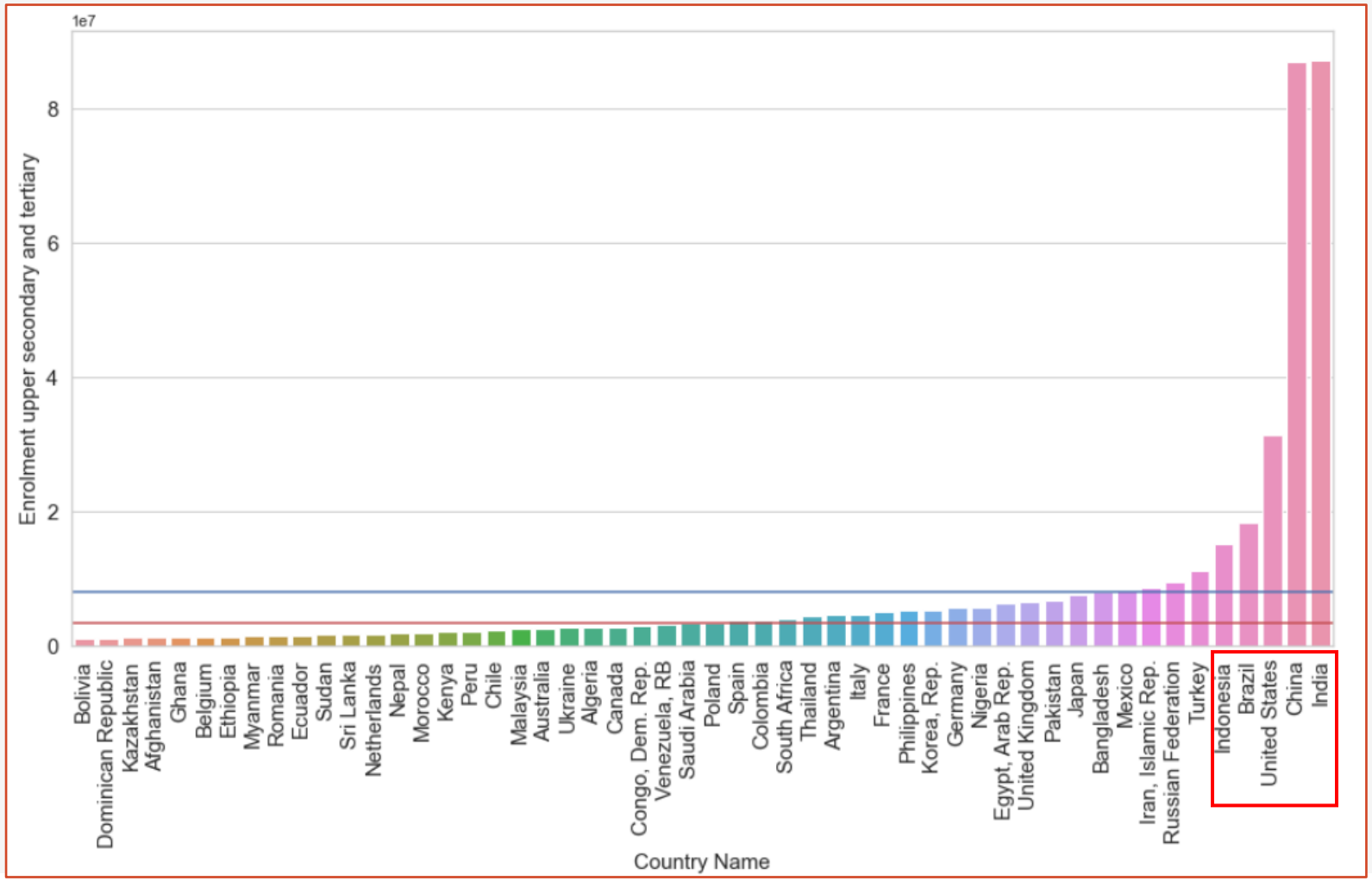
4 Analyse approfondie du jeu de données principal : *Pays*

6) Grandeurs statistiques – Enrolment upper secondary and tertiary

Moyenne ≈ 7 971 685

Médiane = 3 279 286

Ecart-type ≈ 16 799 793



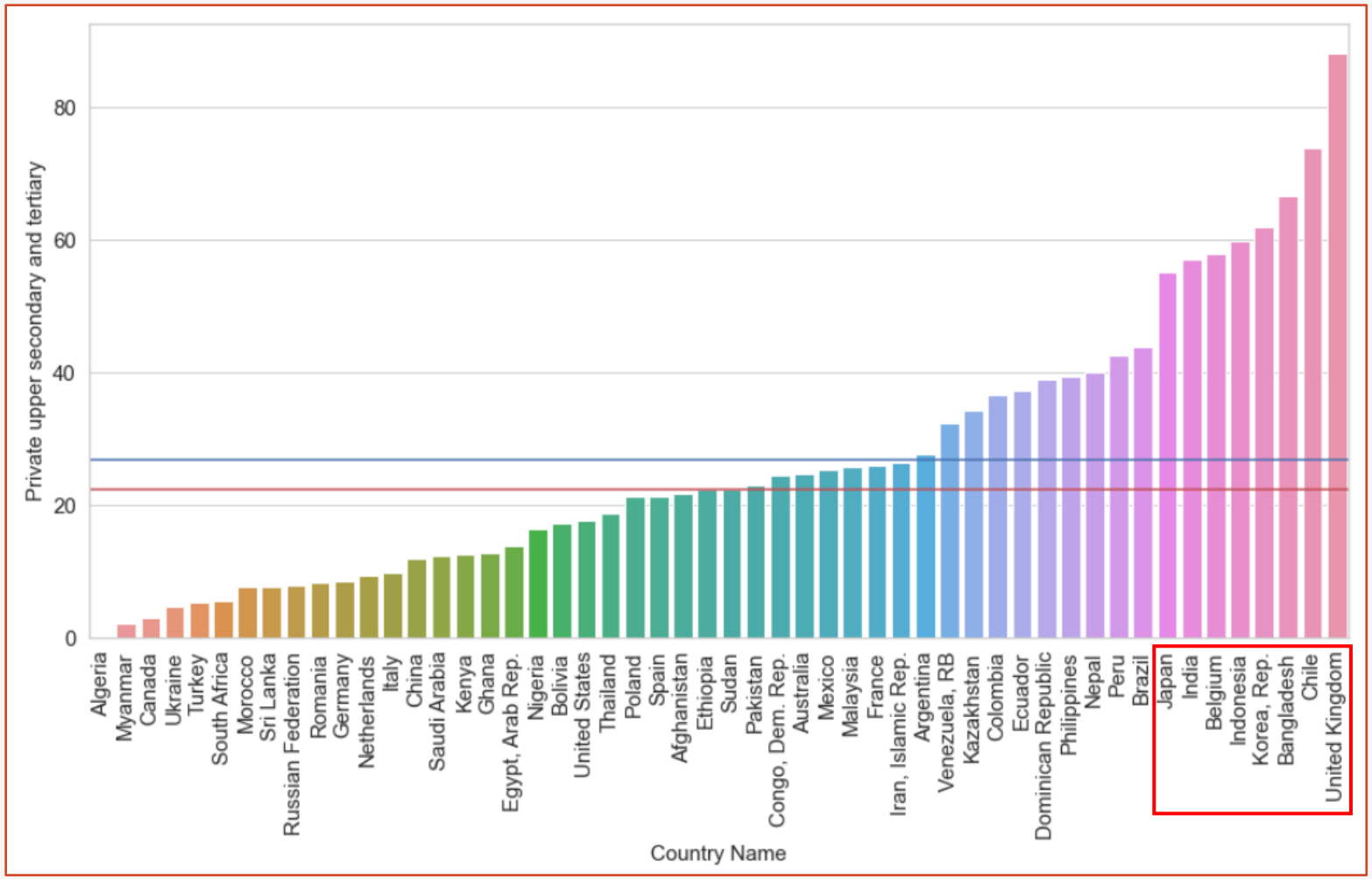
4 Analyse approfondie du jeu de données principal : *Pays*

6) Grandeurs statistiques – Private upper secondary and tertiary

Moyenne ≈ 26,7%

Médiane ≈ 22,3%

Ecart-type ≈ 20,4%



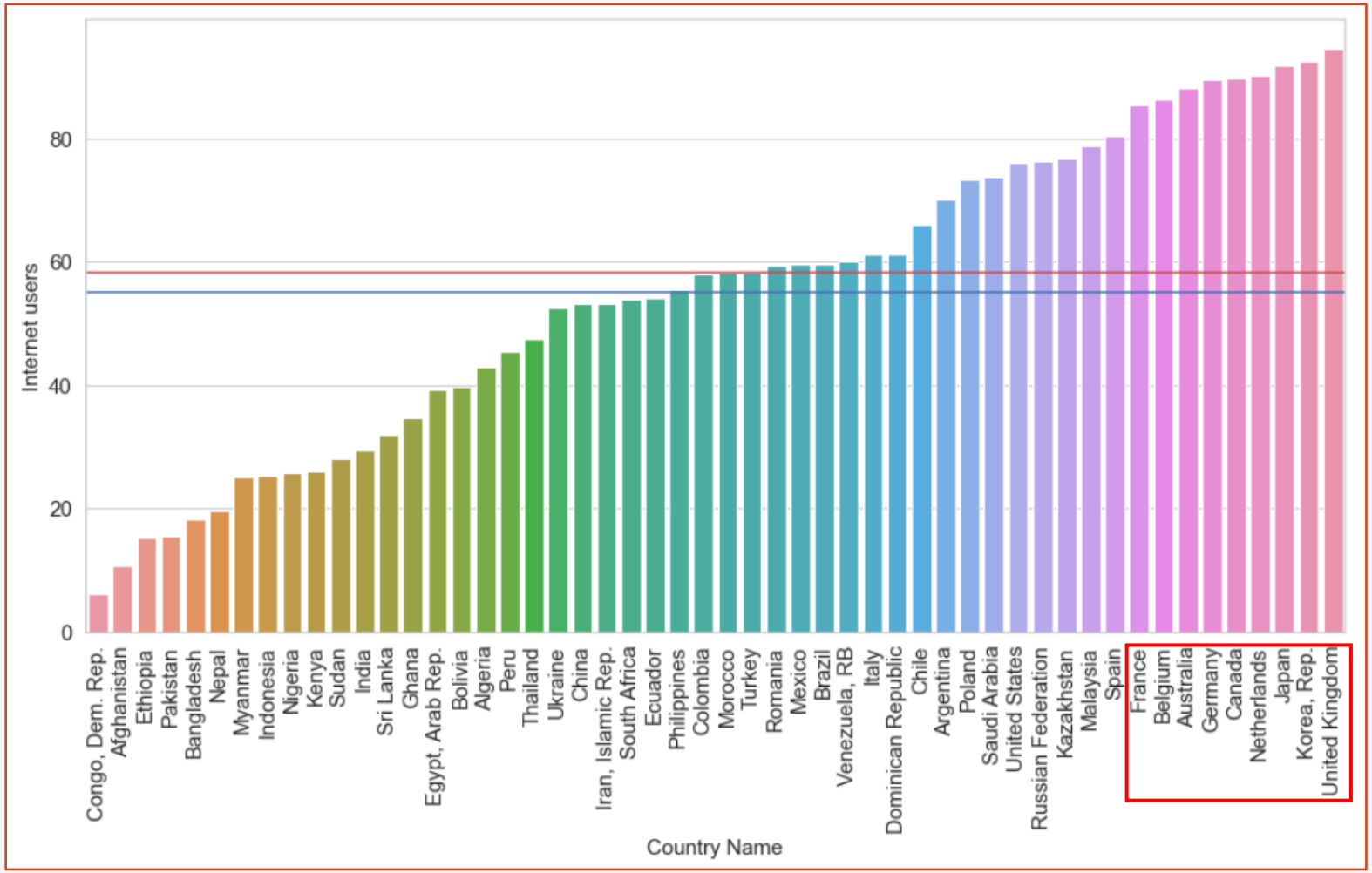
4 Analyse approfondie du jeu de données principal : *Pays*

6) Grandeurs statistiques - Internet Users

Moyenne ≈ 55,1%

Médiane ≈ 58,2%

Ecart-type ≈ 24,9%



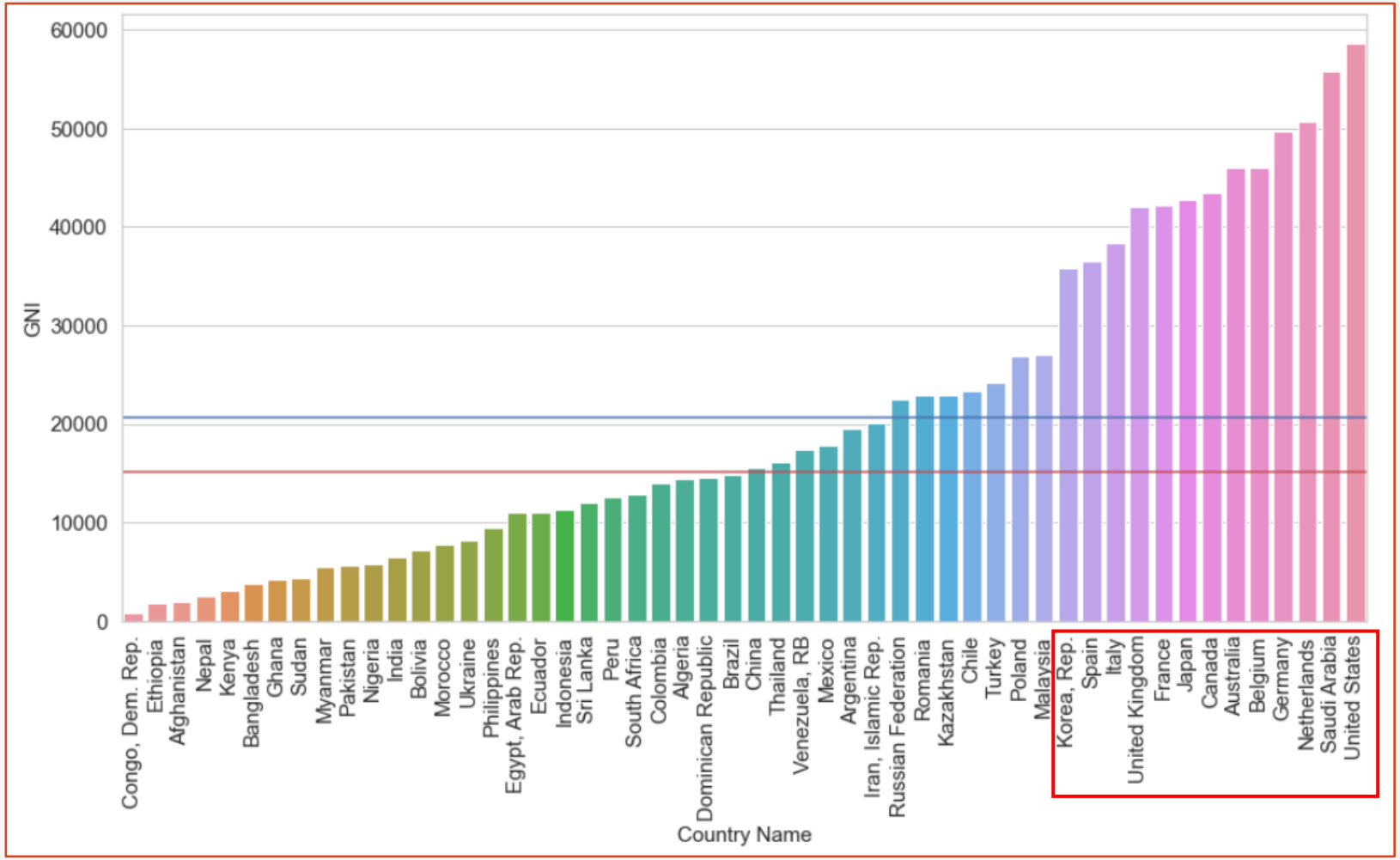
4 Analyse approfondie du jeu de données principal : *Pays*

6) Grandeurs statistiques – GNI

Moyenne ≈ 20588 PPP$

Médiane = 15170 PPP$

Ecart-type ≈ 16166 PPP$



4 Analyse approfondie du jeu de données principal : *Pays*

7) SCORING :

-> Le score de chaque variable de chaque indicateur est ramené entre 0 et 1 (chaque variable est divisée par la valeur maximale de chaque colonne)

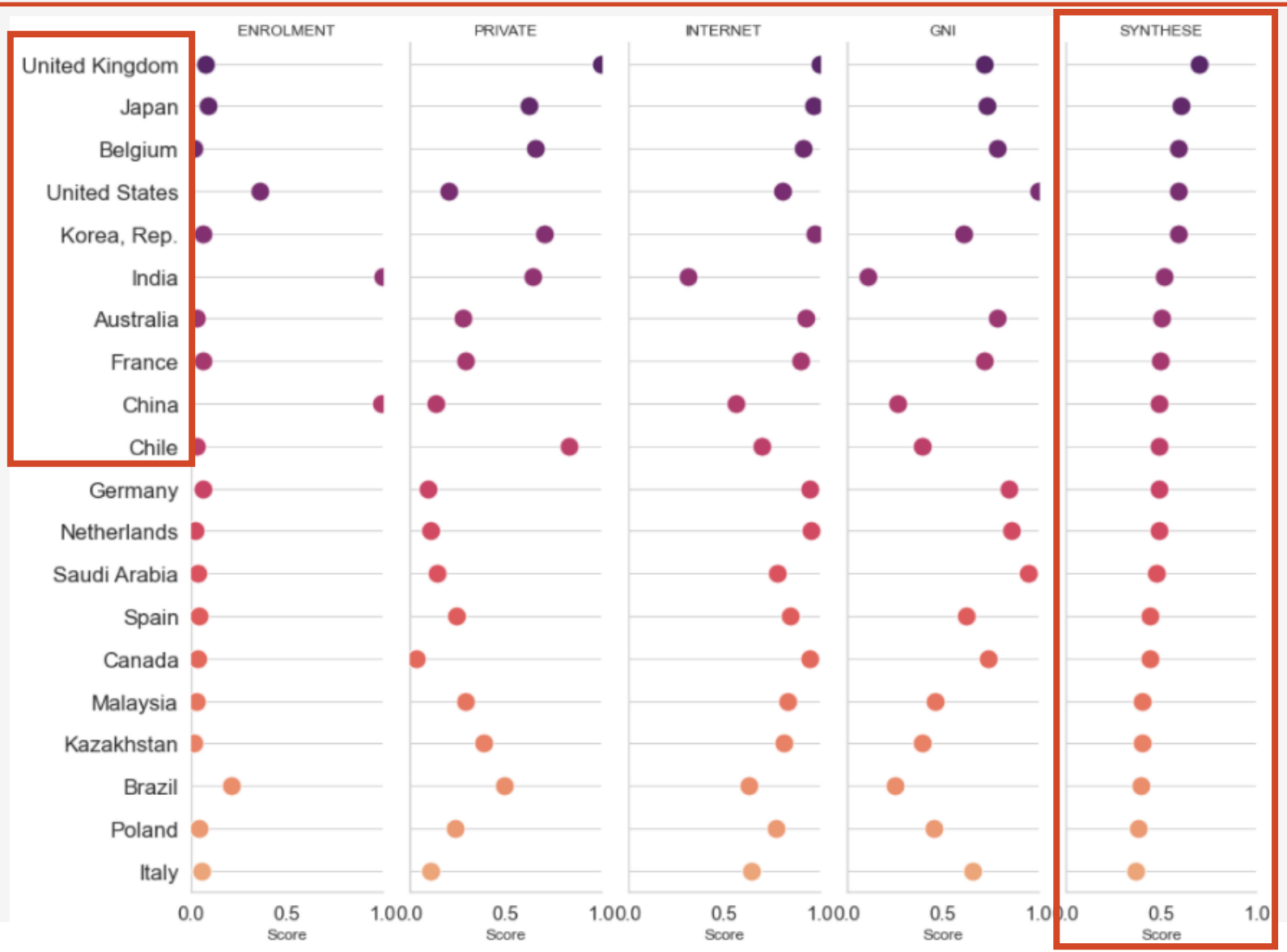
-> Chaque indicateur pèse le même poids (25%) pour le score final

	Country Name	SCORE_ENROLMENT	SCORE_PRIVATE	SCORE_INTERNET	SCORE_GNI	SCORE_SYNT
49	United Kingdom	0.074975	1.000000	1.000000	0.717206	0.698045
24	Japan	0.086395	0.625747	0.970712	0.728961	0.602954
5	Belgium	0.014669	0.655861	0.912854	0.783816	0.591800
50	United States	0.358020	0.199991	0.803757	1.000000	0.590442
27	Korea, Rep.	0.059219	0.702188	0.978272	0.609710	0.587347
20	India	1.000000	0.646637	0.311759	0.110733	0.517282
3	Australia	0.029286	0.278287	0.931025	0.783135	0.505433
17	France	0.057104	0.293531	0.903418	0.717717	0.492943
10	China	0.997029	0.133618	0.561325	0.264055	0.489006
9	Chile	0.025806	0.836286	0.696486	0.396763	0.488835
18	Germany	0.063636	0.095506	0.945886	0.846848	0.487969
33	Netherlands	0.018194	0.106541	0.953946	0.863884	0.485641
41	Saudi Arabia	0.036713	0.140356	0.778162	0.951618	0.476712
43	Spain	0.041524	0.240734	0.850020	0.620954	0.438308
8	Canada	0.031414	0.034102	0.947921	0.739693	0.438283

4 Analyse approfondie du jeu de données principal : *Pays*

7) SCORING :

Top 10

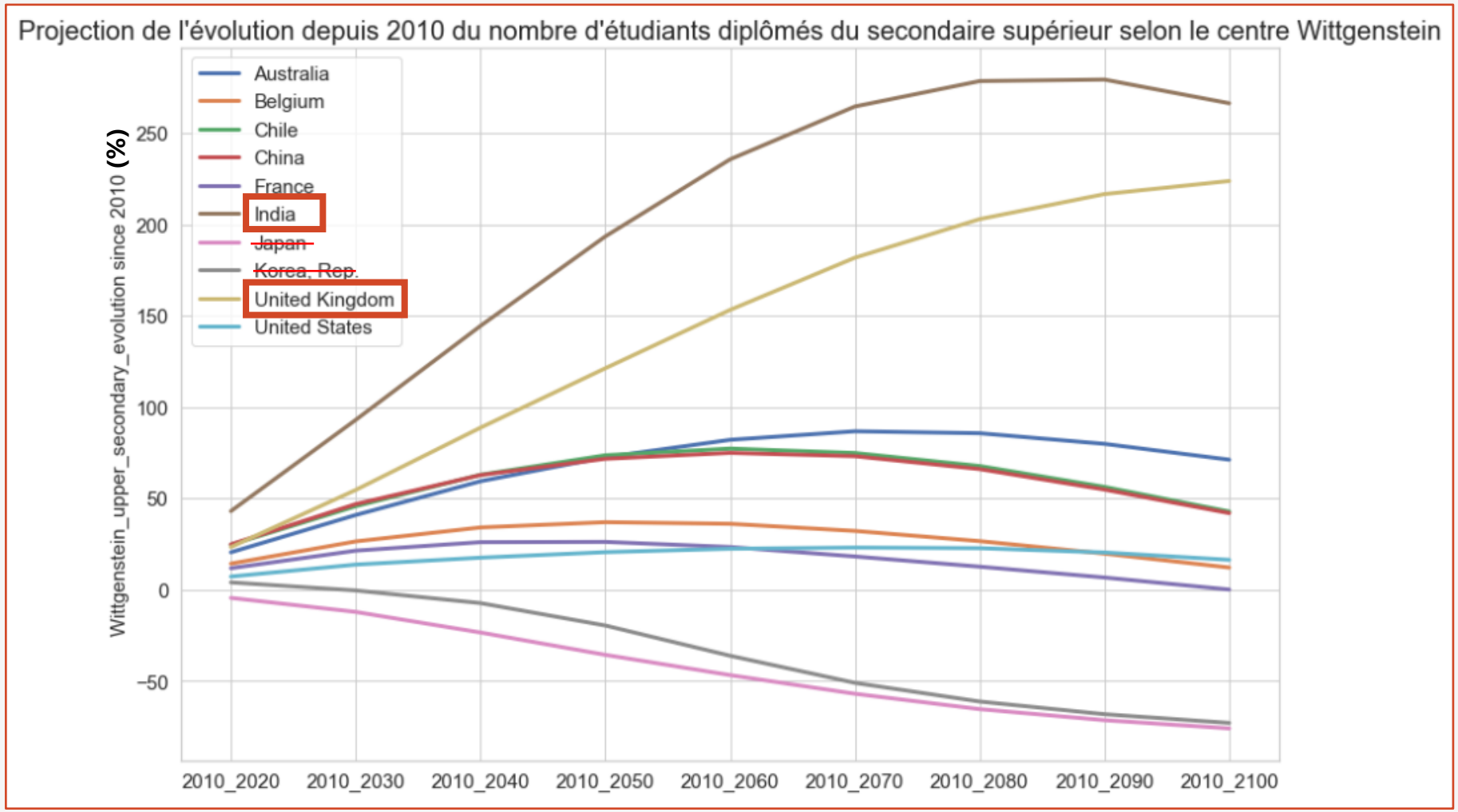


Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

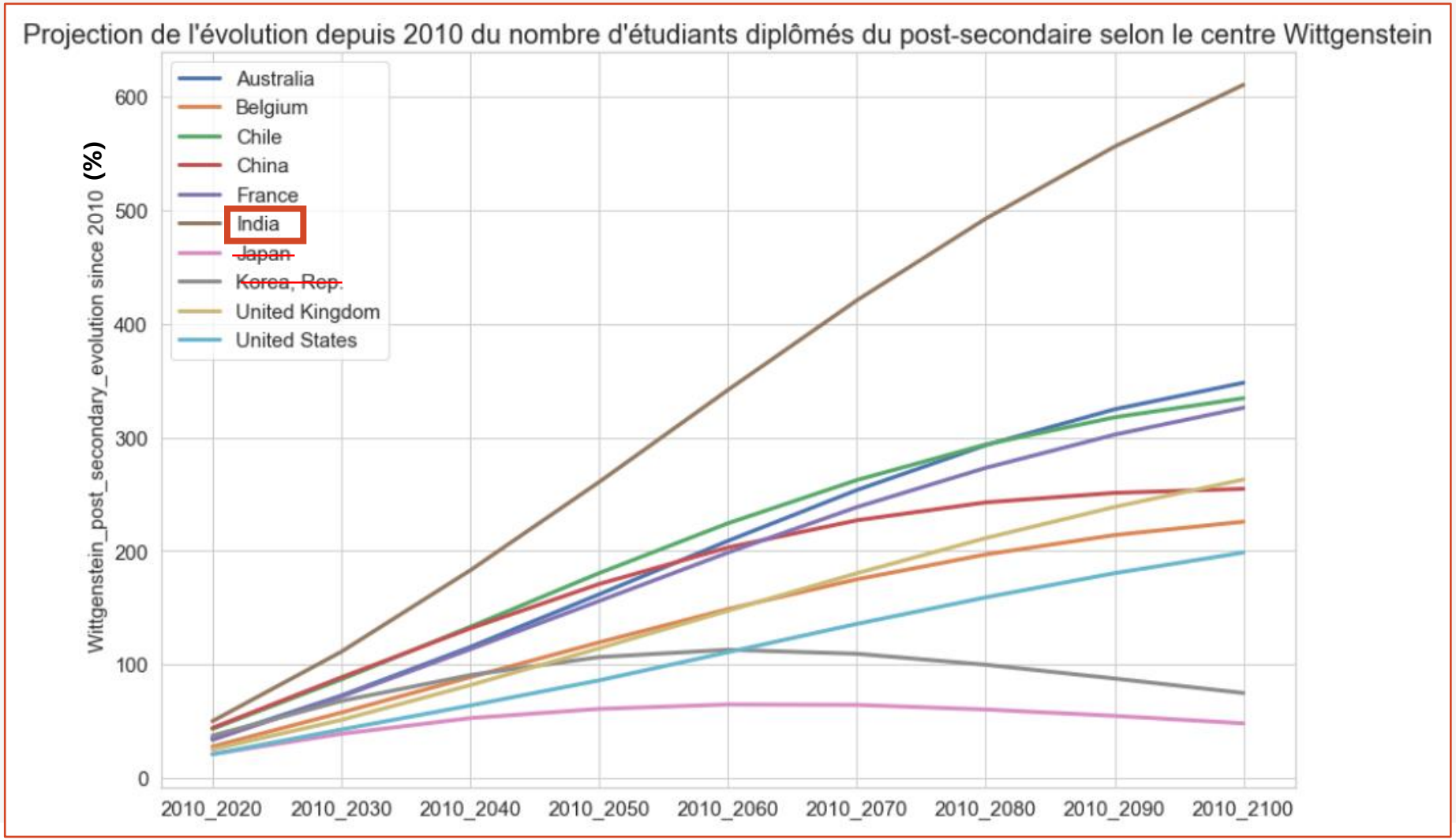
5 Evolution du potentiel de clients : *top 10 des pays*

Indicateur : Wittgenstein Projection: Population in thousands by highest level of educational attainment. **Upper Secondary**. Total



5 Evolution du potentiel de clients : *top 10 des pays*

Indicateur : Wittgenstein Projection: Population in thousands by highest level of educational attainment. **Post Secondary**. Total



Plan de la présentation

1. OBJECTIFS
2. PRESENTATION DES JEUX DE DONNEES
3. CHOIX DES INDICATEURS
4. ANALYSE APPROFONDIE DU JEU DE DONNEES PRINCIPAL
5. EVOLUTION DU POTENTIEL DE CLIENTS
6. CONCLUSIONS ET PERSPECTIVES

6 Conclusions

-> **Pays à fort potentiel de clients ?**

- Top 10 : United Kingdom, Japan, Belgium, United States, Korea Rep., India, Australia, France, China, Chile

-> **Evolution de ce potentiel de clients ?**

- India : +++ ; United Kingdom : ++
- Japan, Korea Rep. : --

-> **Pays dans lequel développer cette nouvelle offre en priorité ?**

- United Kingdom

-> **Pertinence du jeu de données pour répondre aux problématiques de l'entreprise ?**

- Suffisant pour une analyse pré-exploratoire
- Beaucoup de données manquantes – Beaucoup d'indicateurs inexploitables
- Pas de données de projection pour les indicateurs économiques afin d'évaluer l'évolution du potentiel
- Jeu de données pas à jour (données plus récentes sur le site)

Perspectives

-> **Analyser l'évolution du potentiel de clients selon des critères économiques également**

D'après les données d'évolution du GDP de l'Inde (et de la Chine) : bon investissement sur l'avenir

https://www.nextbigfuture.com/2019/01/world-gdp-forecasts-for-2030.html

-> **Discuter du poids de chacun des indicateurs dans le scoring avec l'entreprise pour affiner l'analyse**

Le scoring a été testé sans l'indicateur « GNI » qui a une forte corrélation avec l'indicateur « internet users »

-> pas de changement de top1 mais Inde et Chine remontent dans le classement

-> **Chercher les données des indicateurs manquant pertinents et les ajouter au scoring :**

- Human Capital Index (HCI) : données mises à jour sur la databank

https://databank.worldbank.org/indicator/HD.HCI.OVRL?id=c755d342&report_name=EdStats_Indicators_Report&populartype=series

- Initial Housefolding

Merci pour votre attention