A grayscale photograph of a person wearing large headphones, looking intently at a laptop screen. Their hand is resting on their chin in a thoughtful pose. The background is blurred, showing what appears to be a classroom or office setting.

ANALYSE DE DONNEES DE SYSTEMES EDUCATIFS

Sommaire

I

Rappel de la problématique

II

Environnement technique

III

Analyse pré-exploratoire

IV

Pays ciblés

V

Pertinence du jeu de données



I. RAPPEL DE LA PROBLÉMATIQUE

Rappel de la problématique

Academy est une start up de la EdTech qui propose des contenus de **formation en ligne** (eLearnings) pour un public de niveau **lycée** et **université**.

L'entreprise a un projet **d'expansion à l'international** et aimerait pouvoir **quantifier le potentiel d'un pays en vue de son développement commercial**.

L'objectif de cette présentation est de mener une analyse pré-exploratoire des données sur l'éducation en provenance de la Banque Mondiale afin de conclure si elles permettent d'informer le projet d'expansion.





II. ENVIRONNEMENT TECHNIQUE

Mise en place d'un environnement dédié

Anaconda (conda 4.10.3) `(projet2) C:\Users\milie>`

- Installation de packages, création d'un environnement dédié, lancement de jupyter notebook

Jupyter Notebook (6.4.6) jupyter

- Permet de détailler la démarche et d'exécuter le code directement ligne par ligne

Python (3.8.12)

- Langage de programmation utilisé. Permet d'accéder aux librairies spécifiques et aux fonctions de base (boucles, listes etc) et d'automatiser certaines tâches

Librairies spécifiques au projet

- numpy, pandas, seaborn, matplotlib, nltk, pycharts

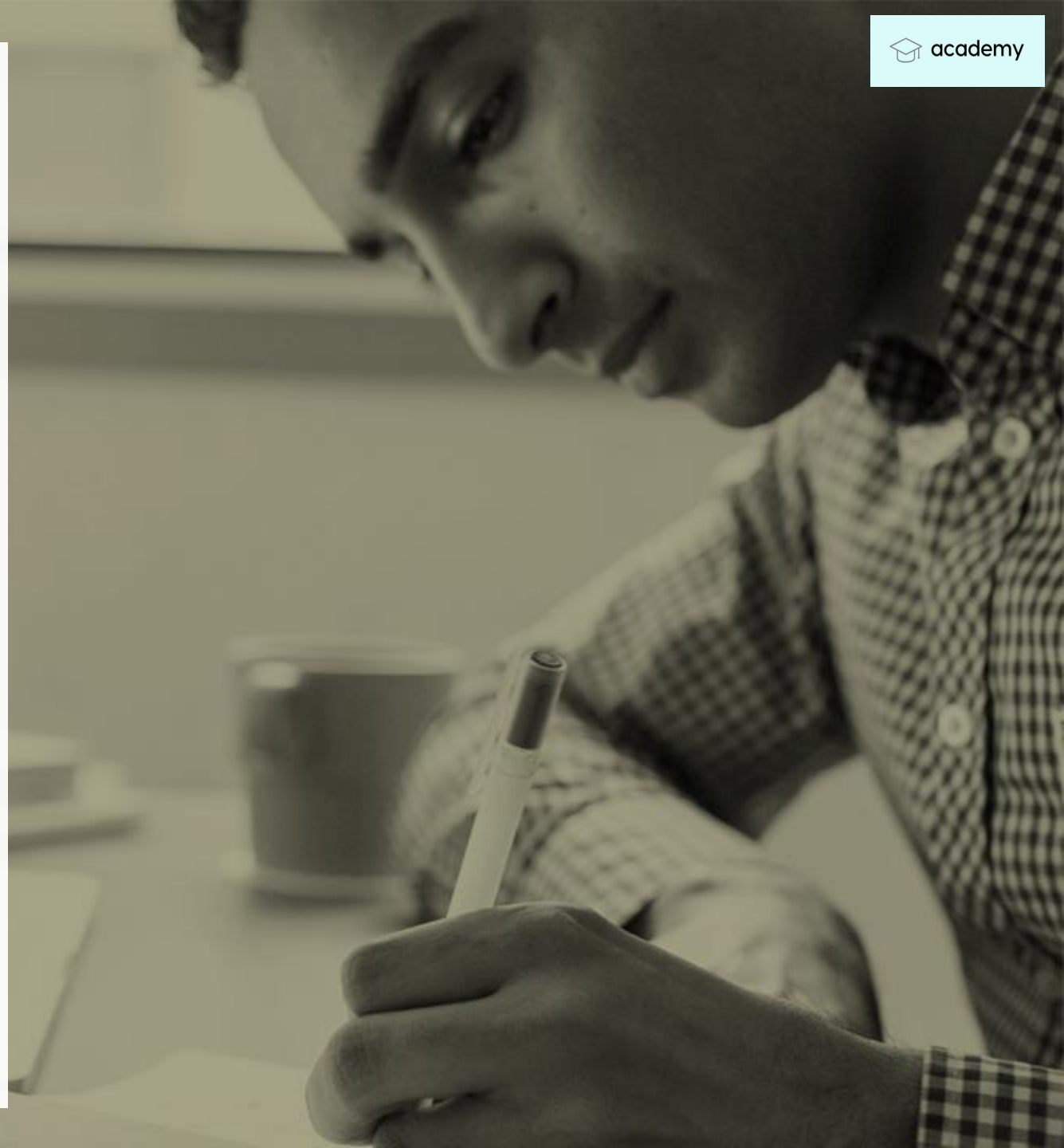




III. ANALYSE PRÉ- EXPLORATOIRE

Méthodologie

- Connaissance des données
- Enrichissement et filtres sur les données
- Sélection des indicateurs pertinents
- Création d'un score moyen pondéré
- Evolution des indicateurs dans le temps
- Validation de la sélection



Connaissance des données

Présentation du jeu de données

EdStatsCountry

Informations générales économiques par pays: région, groupe de revenus, année de mise à jour des données

1 ligne par pays, 241 lignes / 32 colonnes, pas de doublon

De nombreuses valeurs manquantes sur des variables non pertinentes

Variables d'intérêt: Région et Income Group (11% de NaN chacun)

EdStatsCountry-Series

Indicateurs utilisés par identifiant pays et description de la source des données

Plusieurs lignes par pays correspondant à chaque indicateur utilisé, 613 lignes / 4 colonnes, pas de doublon

Aucune valeur manquante (sauf colonne « Unnamed: 3 » composée à 100% de NaN)

211 pays vs 241 pour EdStatsCountry (après vérification ce dernier contient des regroupements de pays: World etc)

Variable d'intérêt: CountryCode

EdStatsSeries

Définitions des indicateurs utilisés dans EdStatsData, affectation des indicateurs dans des « Topics »

1 ligne par indicateur, 3 665 lignes / 21 colonnes, pas de doublon

6 colonnes composées à 100% de NaN et de nombreuses valeurs manquantes sur des variables non pertinentes

Variable d'intérêt: Topics, indicateurs et leur définition

EdStatsFootNote

Mode de calcul des indicateurs par pays et année de référence (1970 à 2050)

643 638 lignes / 5 colonnes (dont une composée à 100% de NaN), pas de doublon

Sur 1 558 indicateurs de EdStatsFootNote, 106 ne sont pas présents dans EdStatsSeries

Variable d'intérêt: Year

EdStatsData

Indicateurs par pays et groupes de pays par année (1970 à 2100)

886 930 lignes / 70 colonnes (dont une composée à 100% de NaN), pas de doublon

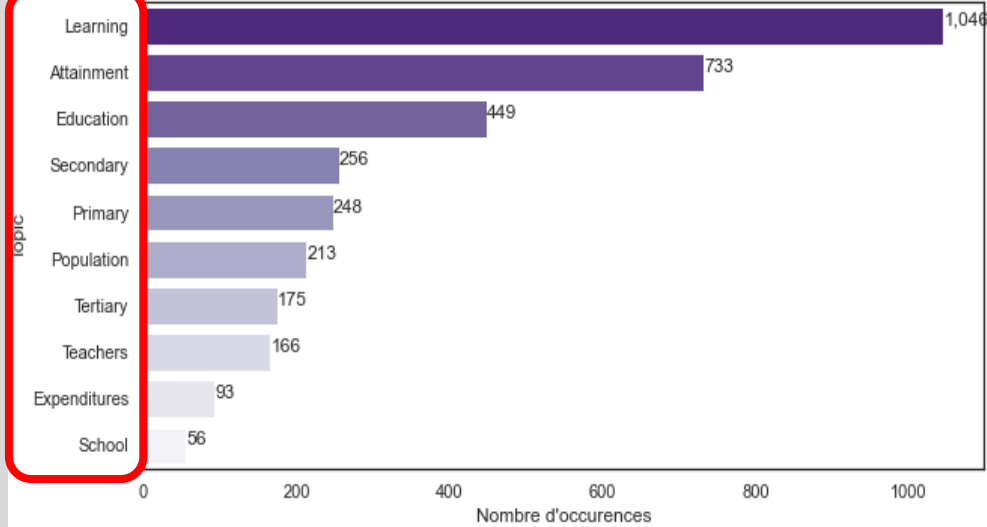
De nombreuses données manquantes par année

242 pays vs 211 pour EdStatsCountry-Series

Variable d'intérêt: Toutes

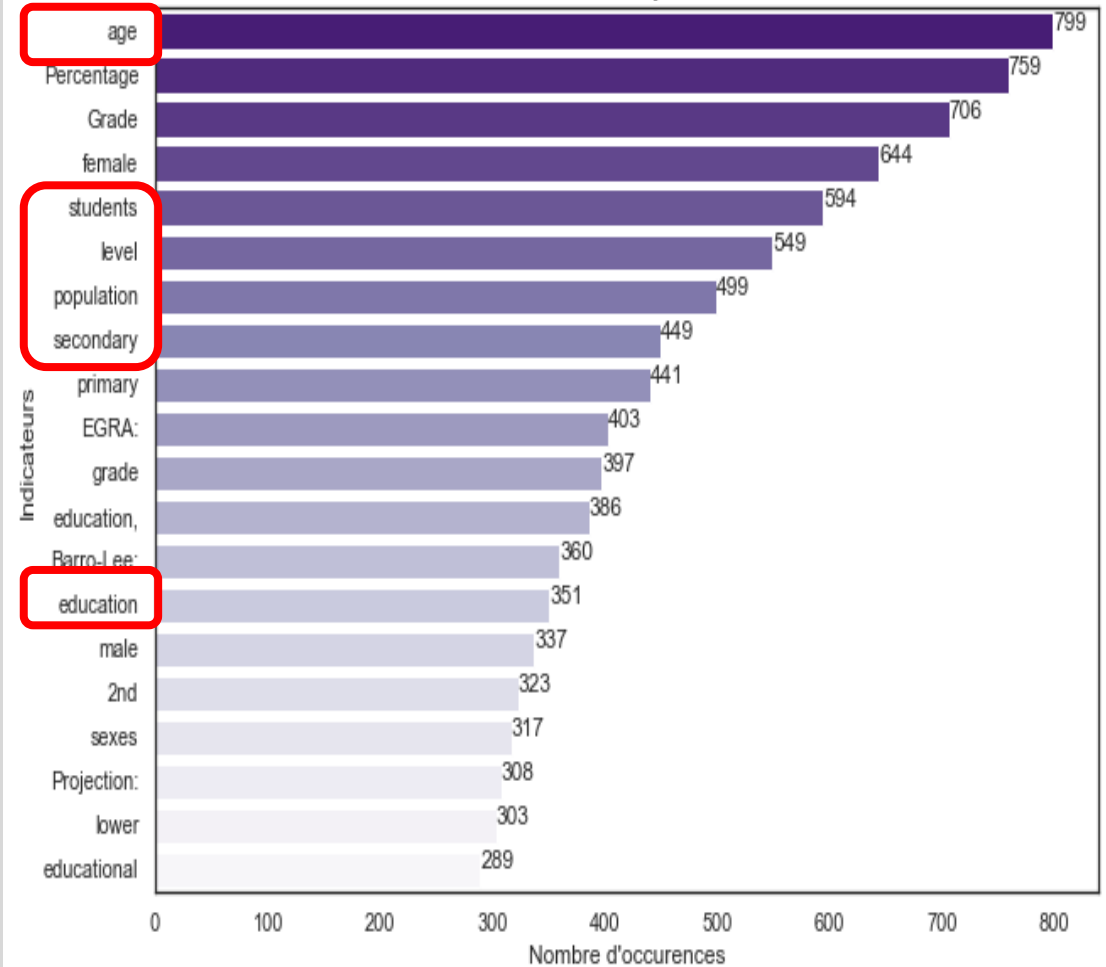
Vue d'ensemble

TOP 10 des Topics les plus utilisées

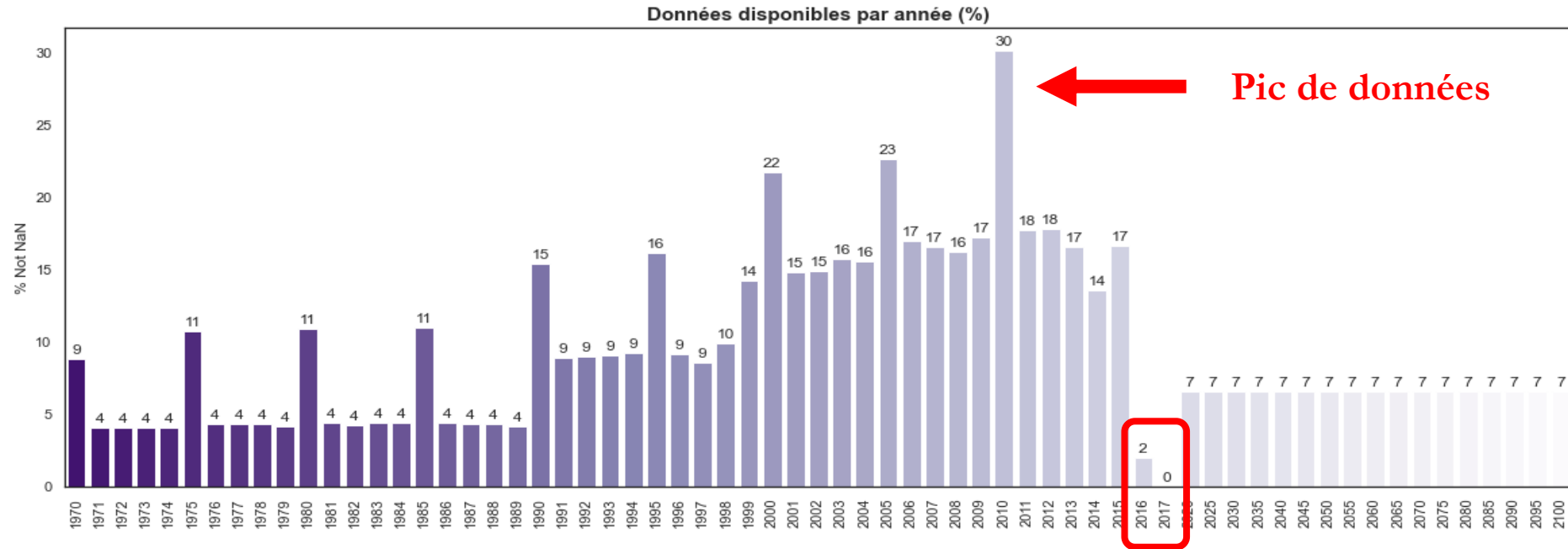


- 242 pays réduits à **211** (suppression des regroupements)
- **3 665** indicateurs uniques

TOP 20 des mots revenant le plus dans les indicateurs



Données disponibles par année

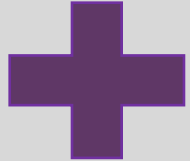


- Données de 1970 à 2100
- Pics de données tous les 5 ans
- Année 2010 la plus fournie
- Aucune donnée pour l'année 2017 (2% pour 2016)
- Années 1990 à 2015 contiennent le plus de données
- Même nombre de données à partir de 2020

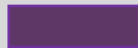
Enrichissement et filtres sur les données

Intégration et suppression de données

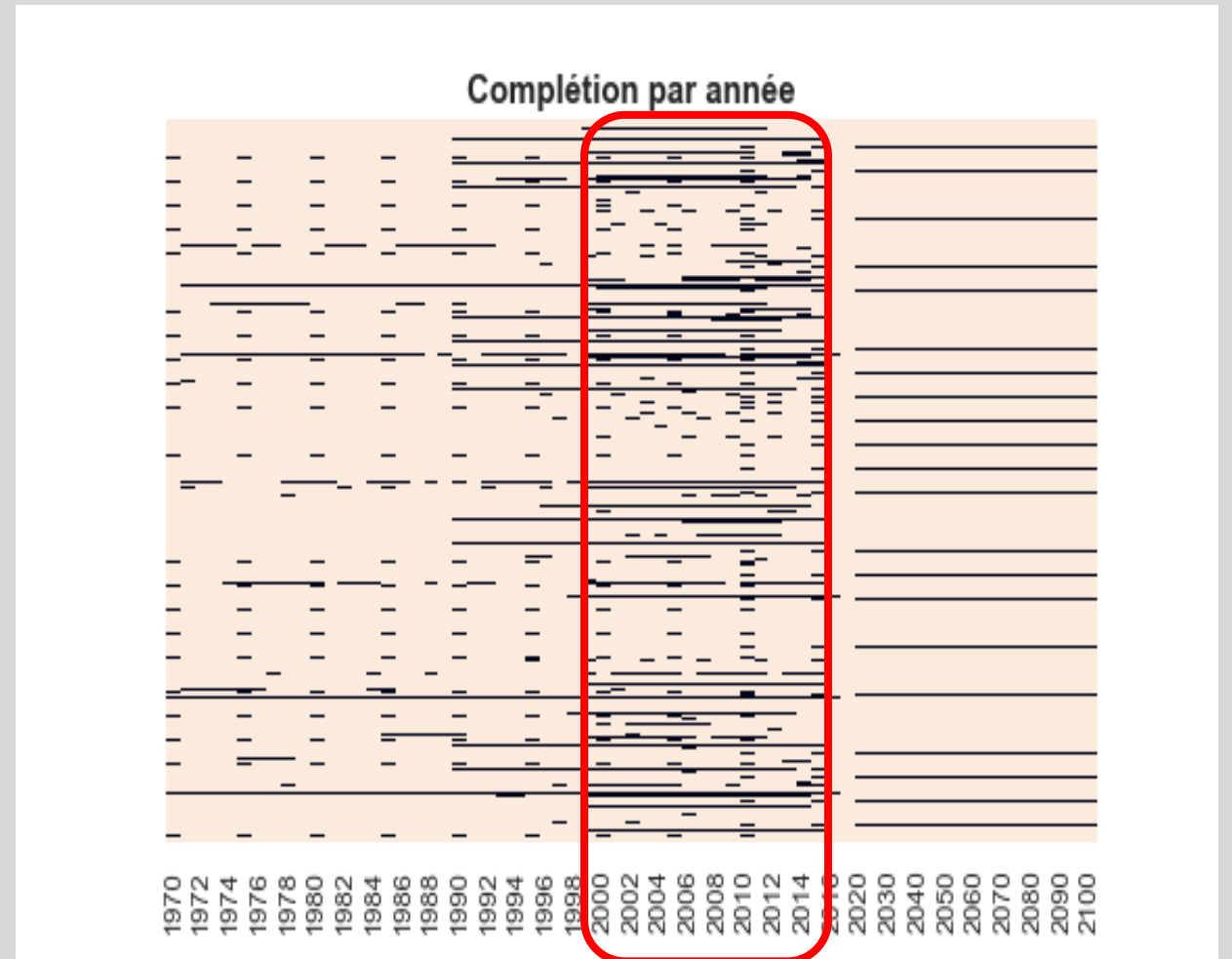
A partir du fichier EdStatsData:



- Intégration des variables **Region** et **Income Group** issues de EdStatsCountry
- Intégration de la variable **Topic** issue de EdStatsSeries



- Filtre sur les pays présents dans EdStatsCountry-Series (**211 pays** sans regroupement)
- Filtre sur les **années supérieures aux années 2000** (tous les 5 ans)
- Filtre sur les **mots** contenus dans les indicateurs avec la plus forte **occurrence**



Sélection des indicateurs pertinents

Recherche des indicateurs pertinents

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

Population
de lycéens?

Population
d'étudiants
à l'université

Richesse
du pays?

Accès au
numérique

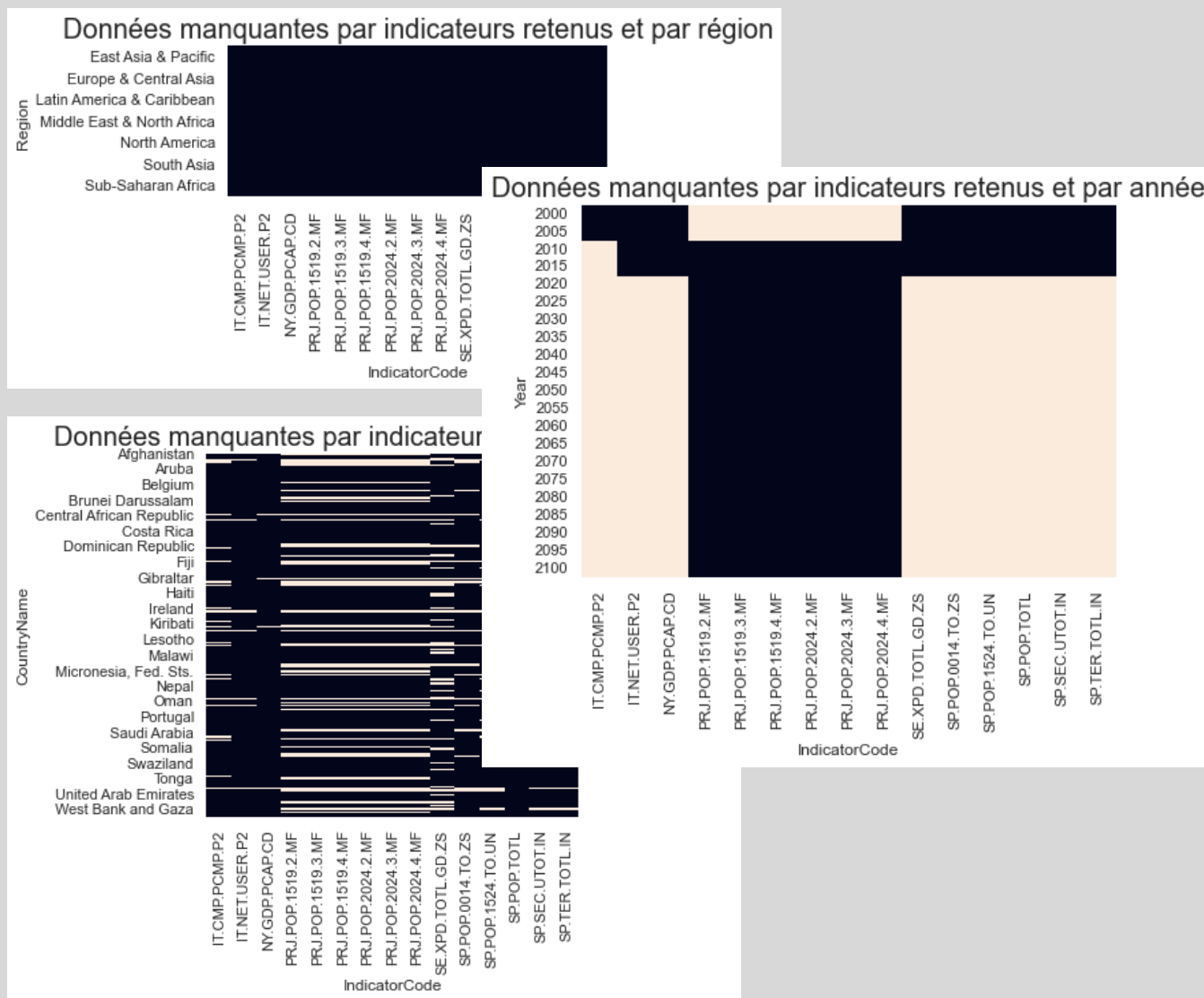
Niveau
d'anglais,
langues
parlées?

Dépenses
dans
l'éducation?

Indicateurs retenus et validation de la sélection

Indicateurs retenus:

- population en âge de faire des **études secondaires supérieures**
- population en âge de faire des **études tertiaires**
- Utilisateurs d'**internet** (%)
- **Ordinateur personnel** (%)
- **PIB** par habitant (US\$)
- **Dépenses du Gouvernement dans l'éducation** (% PIB)
- **Population totale**
- Population, **âges 0-14** (% du total)
- Population, **âges 15-24**, total
- Projections: population **15-24** ans en milliers avec **au moins un niveau secondaire**



Exemple de statistiques par groupe de revenus, pays et indicateur

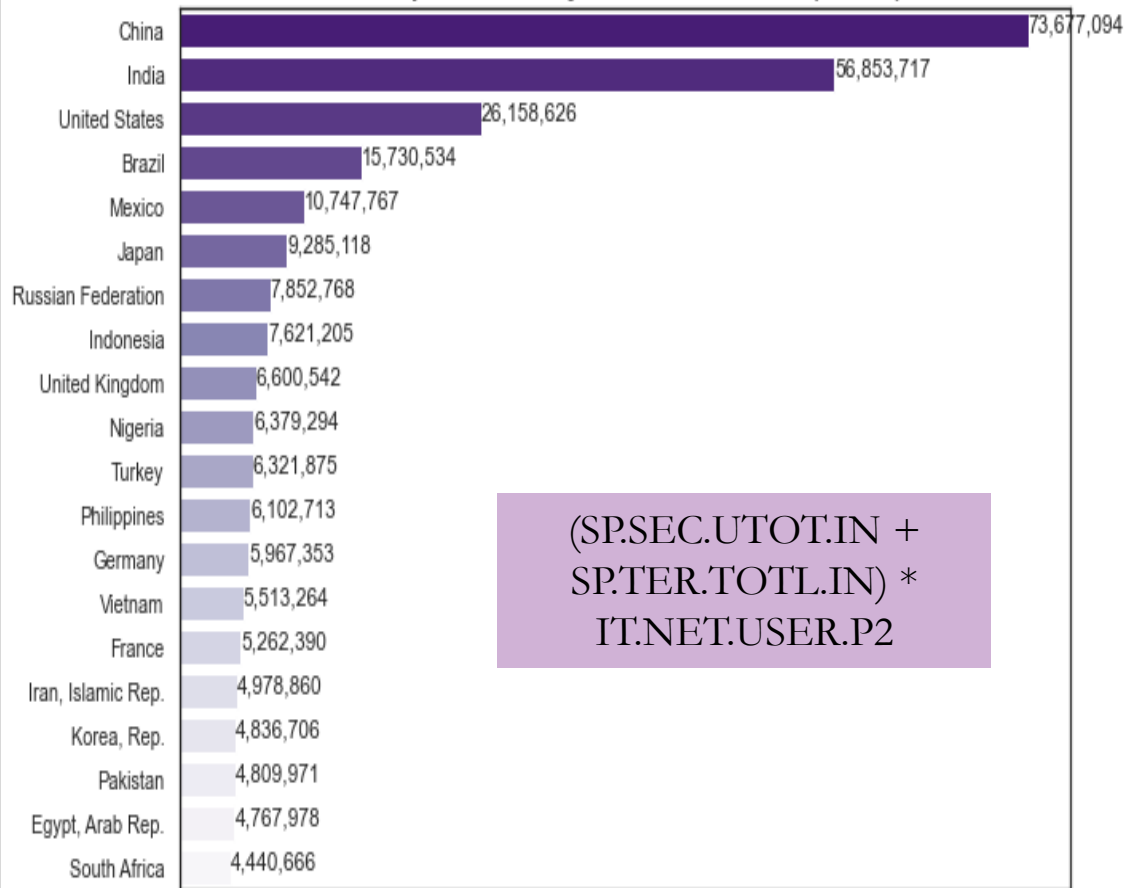
IndicatorCode		IT.CMP.PCMP.P2					IT.NET.USER.P2					NY.GDP.PCAP.CD				
IndicatorName		Personal computers (per 100 people)					Internet users (per 100 people)					GDP per capita (current US\$)				
		mean	median	std	min	max	mean	median	std	min	max	mean	median	std	min	max
IncomeGroup	CountryName															
High income: OECD	Australia	46.73	46.73	NaN	46.73	46.73	67.58	69.50	16.47	46.76	84.56	41033.94	42945.40	16142.86	21690.92	56554.04
	Austria	48.40	48.40	17.31	36.16	60.64	62.71	66.59	22.12	33.73	83.94	38520.30	41329.36	9953.07	24564.46	46858.04
	Belgium	30.01	30.01	10.77	22.40	37.63	61.33	65.41	24.48	29.43	85.05	36232.56	38671.32	9196.76	23207.41	44380.18
	Canada	64.88	64.88	32.32	42.02	87.73	72.93	75.98	15.97	51.30	88.47	37769.23	39752.64	10216.24	24124.17	47447.48
	Chile	11.63	11.63	3.46	9.19	14.08	39.27	38.09	20.32	16.60	64.29	9807.52	10237.74	4125.60	5101.37	13653.23
	Czech Republic	19.78	19.78	10.73	12.19	27.37	47.38	52.05	30.66	9.78	75.67	14220.37	15530.90	6098.83	6011.62	19808.07
	Denmark	60.10	60.10	13.47	50.58	69.62	76.74	85.73	25.66	39.17	96.33	47649.45	50906.41	11886.87	30743.56	58041.41
	Estonia	32.58	32.58	23.30	16.10	49.05	63.13	67.78	25.54	28.58	88.41	11550.71	12488.46	5726.65	4070.03	17155.87
	Finland	44.82	44.82	7.38	39.60	50.03	71.26	80.45	23.39	37.25	86.89	37961.10	40694.37	9604.14	24253.25	46202.42
	France	43.61	43.61	18.88	30.26	56.96	54.79	60.08	32.56	14.31	84.69	33643.87	35703.25	7844.90	22465.64	40703.34

Construction de la table des indicateurs

CountryName	Students_Sec	Students_Ter	Students	%_Internet	Prospects_Internet	%_Ordinateur	Prospects_Ordinateurs	GDP_per_capita	Gov_expenditures_%_GDP	Pop_15-24_2040
China	46531520.00	99943816.00	146475336.00	50.30	73677094.01	4.82	7058319.56	8069.21	3.83	121684.58
India	99198160.00	119469984.00	218668144.00	26.00	56853717.44	1.51	3298034.62	1596.47	3.42	198413.15
United States	12321350.00	22765372.00	35086722.00	74.55	26158625.75	77.27	27112367.78	56469.01	5.43	43833.77
Brazil	10596253.00	16372865.00	26969118.00	58.33	15730534.13	16.12	4346530.01	8757.21	5.64	25291.14
Mexico	7180821.00	11533390.00	18714211.00	57.43	10747766.56	12.64	2366067.31	9152.87	5.18	18430.18
Japan	3591159.00	6605765.00	10196924.00	91.06	9285117.95	31.82	3244465.80	34474.14	3.64	8902.36
Russian Federation	2497011.00	8200126.00	10697137.00	73.41	7852768.27	12.09	1293172.71	9329.30	3.77	14738.71
Indonesia	13619948.00	21059620.00	34679568.00	21.98	7621205.35	1.46	507492.53	3336.11	3.59	33487.58
United Kingdom	2977214.00	4197265.00	7174479.00	92.00	6600542.20	75.73	5433270.09	44305.55	5.68	8181.27

Visualisations des indicateurs retenus (valeur la plus récente non nulle)

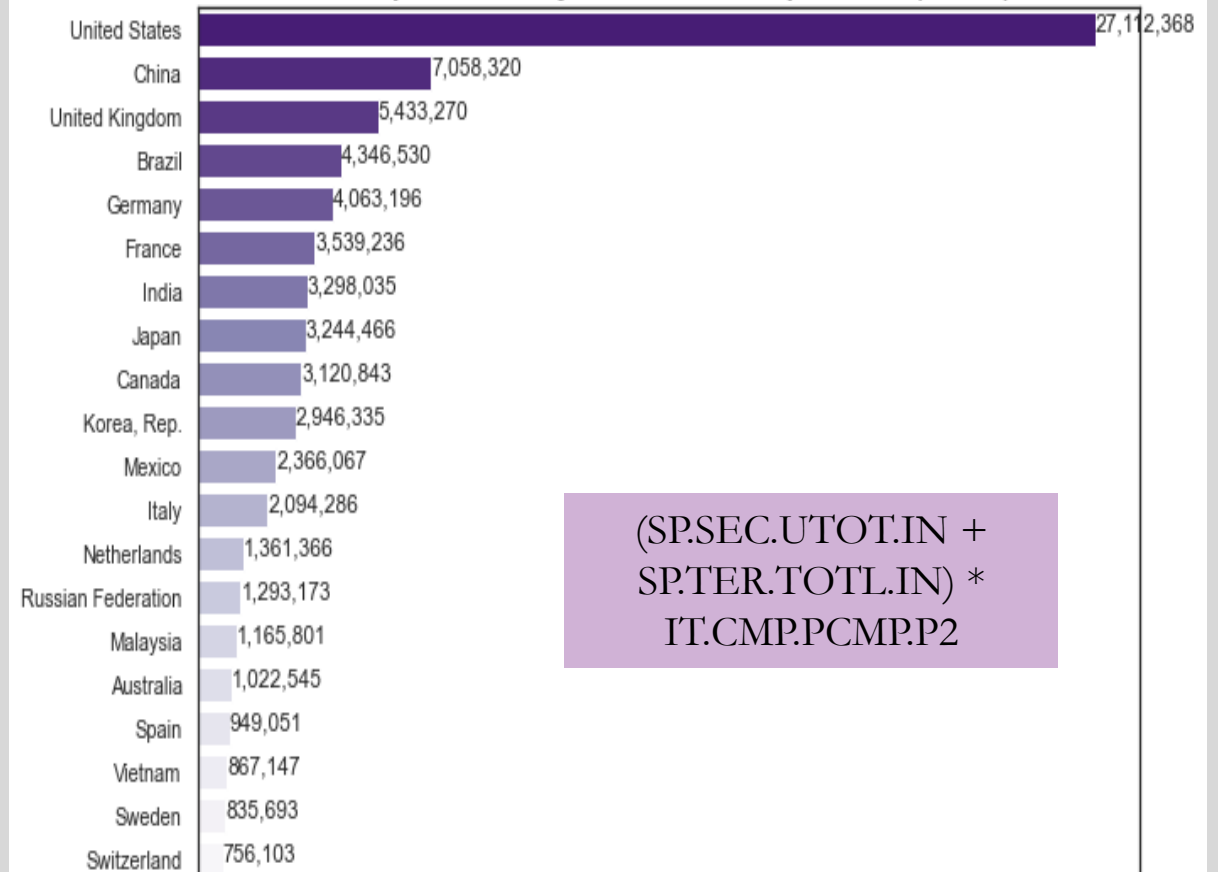
Personnes ayant l'âge de faire des études secondaires et supérieures et ayant accès à internet (TOP20)



$(SP.SEC.UTOT.IN + SP.TER.TOTL.IN) * IT.NET.USER.P2$

Population

Personnes ayant l'âge de faire des études secondaires et supérieures et ayant un ordinateur personnel (TOP20)

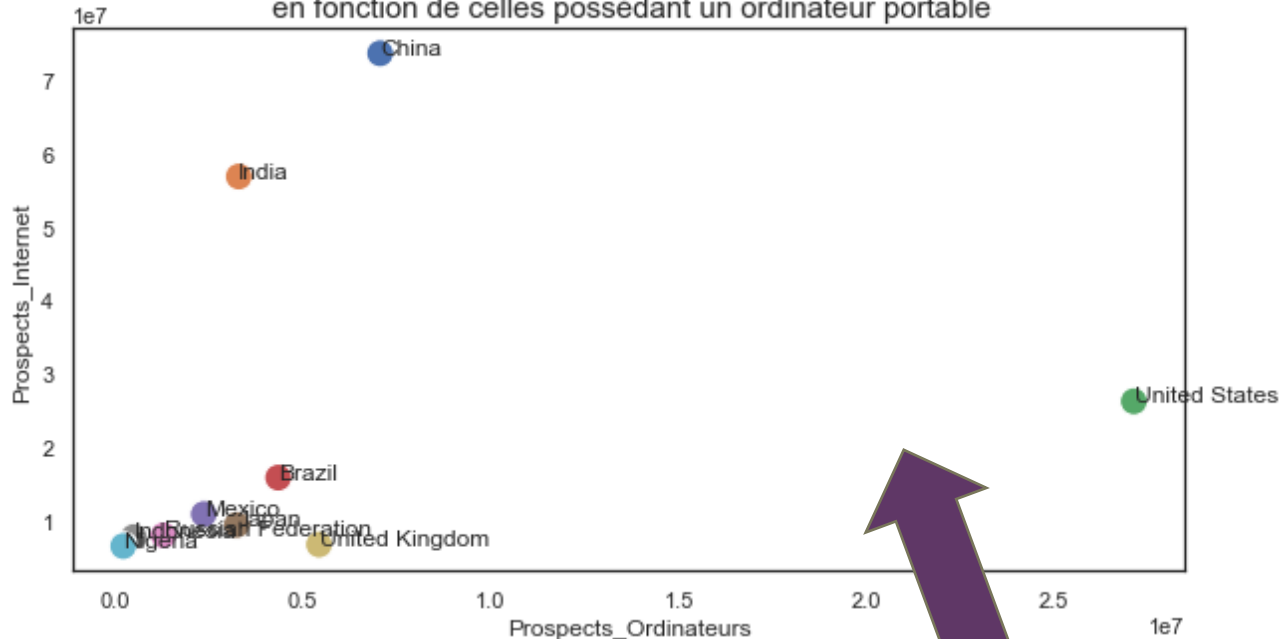


$(SP.SEC.UTOT.IN + SP.TER.TOTL.IN) * IT.CMP.PCMP.P2$

Prospects

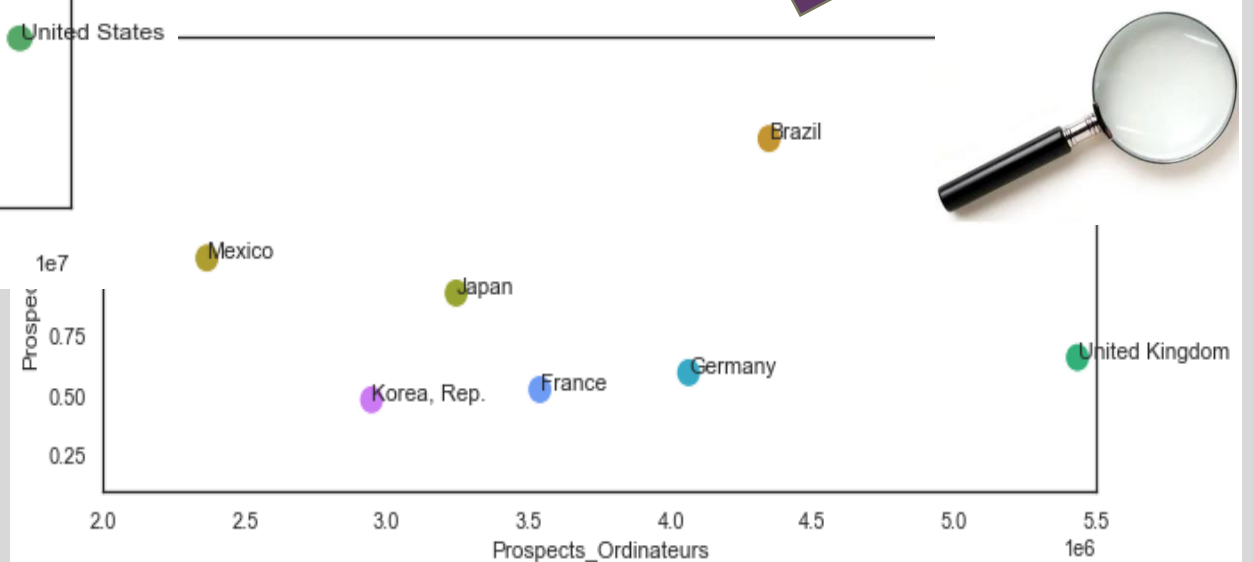
Visualisations des indicateurs retenus (valeur la plus récente non nulle)

Personnes en âge de faire des études secondaires et supérieures avec accès à internet en fonction de celles possédant un ordinateur portable



On remarque que les **Etats Unis**, **l'Inde** et la **Chine** sont les pays combinant le plus de clients potentiels: âge de faire des études secondaires ou supérieures, accès à internet et possédant un ordinateur personnel.

Le **Brésil**, le **Royaume Uni**, le **Mexique**, le **Japon**, **l'Allemagne**, la **France** et la **Corée** sont des pays qui constituent également un bon vivier de clients potentiels.

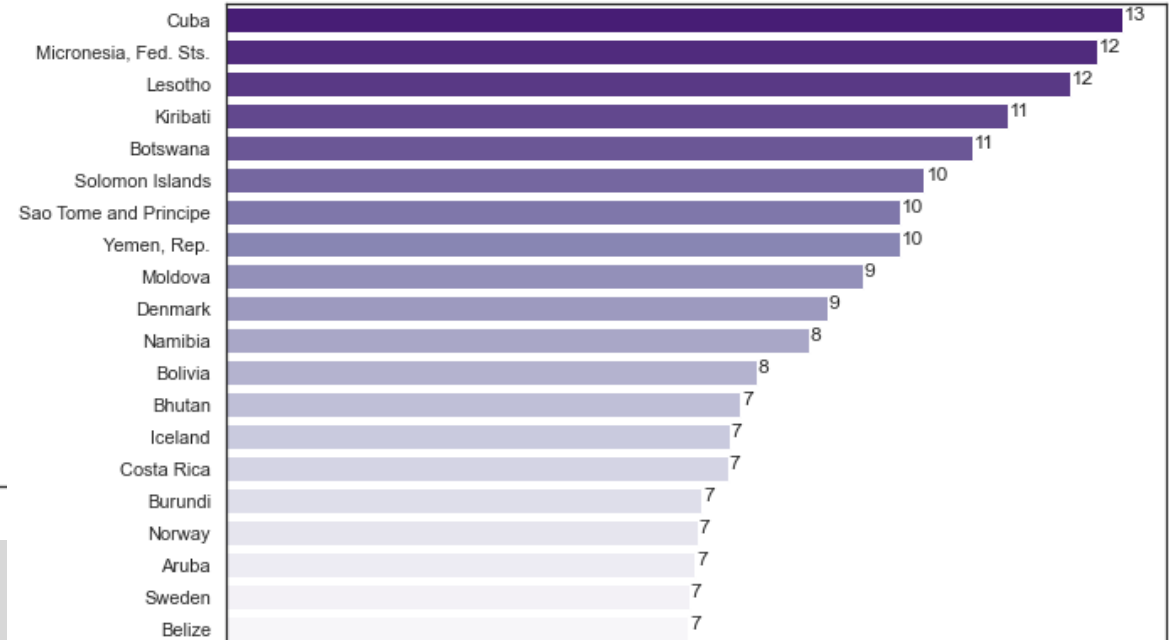


Visualisations des indicateurs retenus (valeur la plus récente non nulle)

SE.XPD.TOTL.GD.ZS

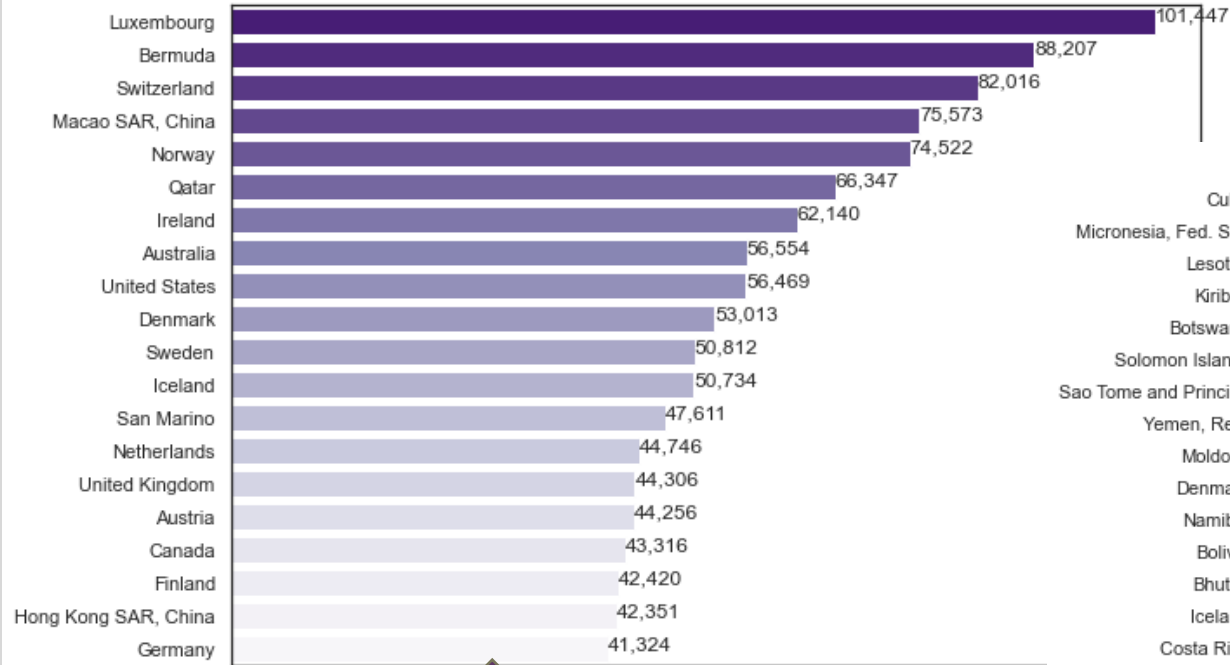


Dépenses du gouvernement dans l'éducation (% du PIB)



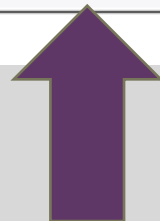
% du PIB

PIB par habitant



GDP_per_capita

NY.GDP.PCAP.CD



Création d'un score moyen pondéré

Score paramétrable pour aider à la décision

Création de 5
scores et
d'un score
moyen
pondéré

```
def normalize(dataframe, col_name):  
    return dataframe[col_name] / dataframe[col_name].max()
```

```
def score_pays(dataframe,  
               pop_min_prospects_internet,  
               poids_score_internet,  
               poids_score_ordi,  
               poids_score_pib,  
               poids_score_gov_expenses,  
               poids_score_pop1524_2040,  
               TOP):
```

```
''' Fonction qui permet de sélectionner la population en âge de faire des études secondaires et tertiaires et  
ayant accès à internet minimale, de recalculer les scores en divisant chaque valeur par la valeur maximale de la  
sélection, et de pouvoir affecter des poids différents pour chaque score afin de lui donner plus ou moins  
d'importance. Le nombre de pays retenus est également paramétrable. La fonction retourne un classement  
des pays par ordre décroissant en fonction du score moyen.  
'''
```

```
classement_pays_filtered = dataframe[dataframe.Prospects_Internet >= pop_min_prospects_internet]  
classement_pays_filtered['SCORE_SYNT'] = [0]*len(classement_pays_filtered)  
poids_liste = [poids_score_internet, poids_score_ordi, poids_score_pib, poids_score_gov_expenses,  
               poids_score_pop1524_2040]
```

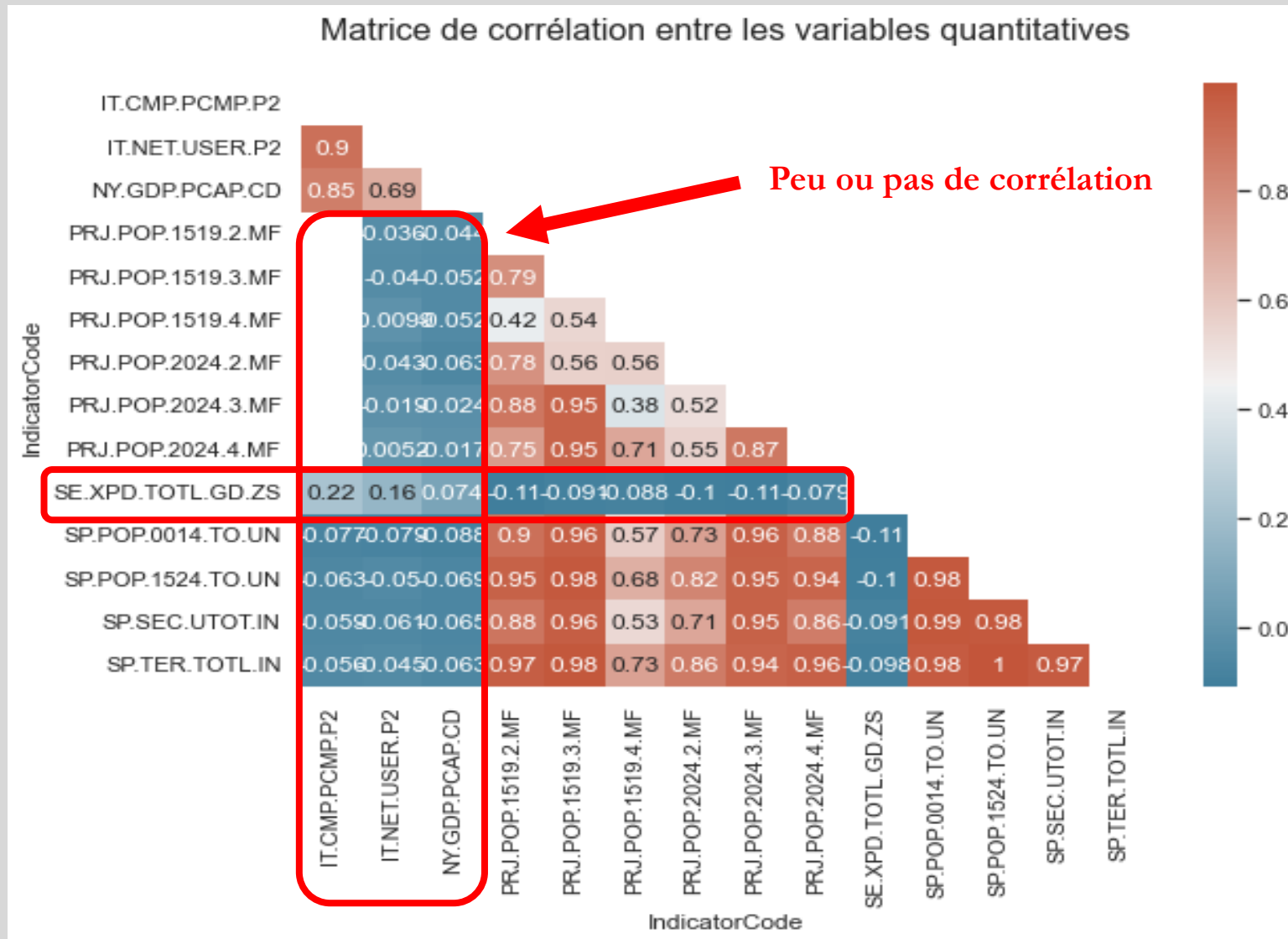
```
for i, c in enumerate(classement_pays_filtered.columns[1:-1]):  
    classement_pays_filtered[f'SCORE_{str(c)}'] = normalize(classement_pays_filtered, c)  
    classement_pays_filtered['SCORE_SYNT'] += classement_pays_filtered[f'SCORE_{str(c)}']*poids_liste[i]
```

```
classement_pays_filtered['SCORE_SYNT'] = classement_pays_filtered['SCORE_SYNT']/len(classement_pays_filtered.columns[1:-1])  
  
new_classement_pays = classement_pays_filtered.drop(columns = ["Prospects_Internet", "Prospects_Ordinateurs",  
                                                             "GDP_per_capita", "Gov_expenditures_%_GDP", "Pop_15-24_2040"])
```

```
return new_classement_pays.sort_values(by = 'SCORE_SYNT', ascending = False).reset_index(drop = True).head(TOP)
```

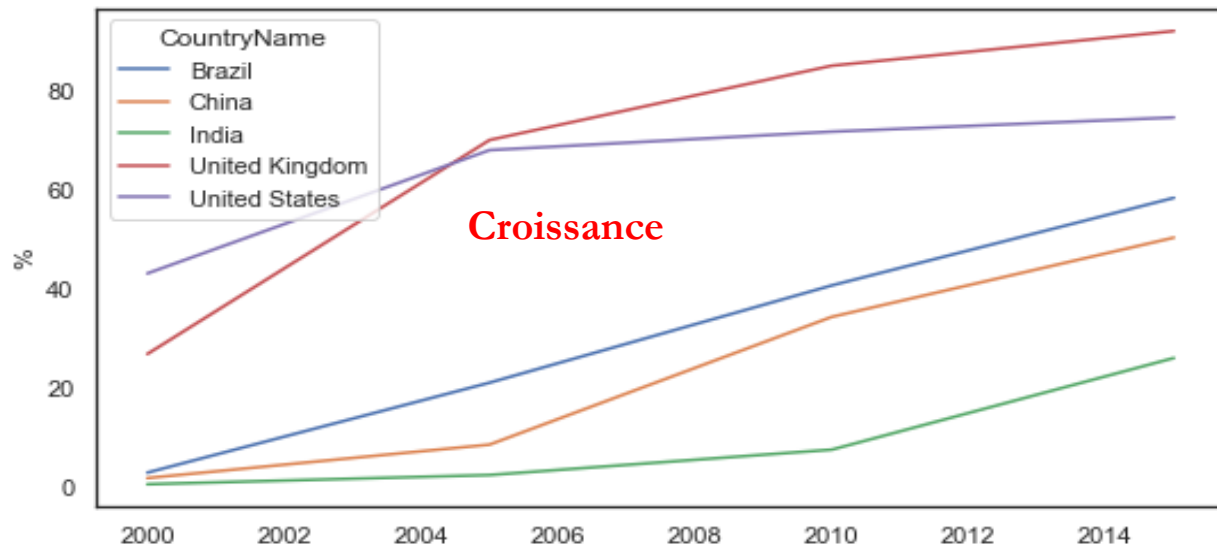

Evolution des indicateurs dans le temps

Corrélation des indicateurs

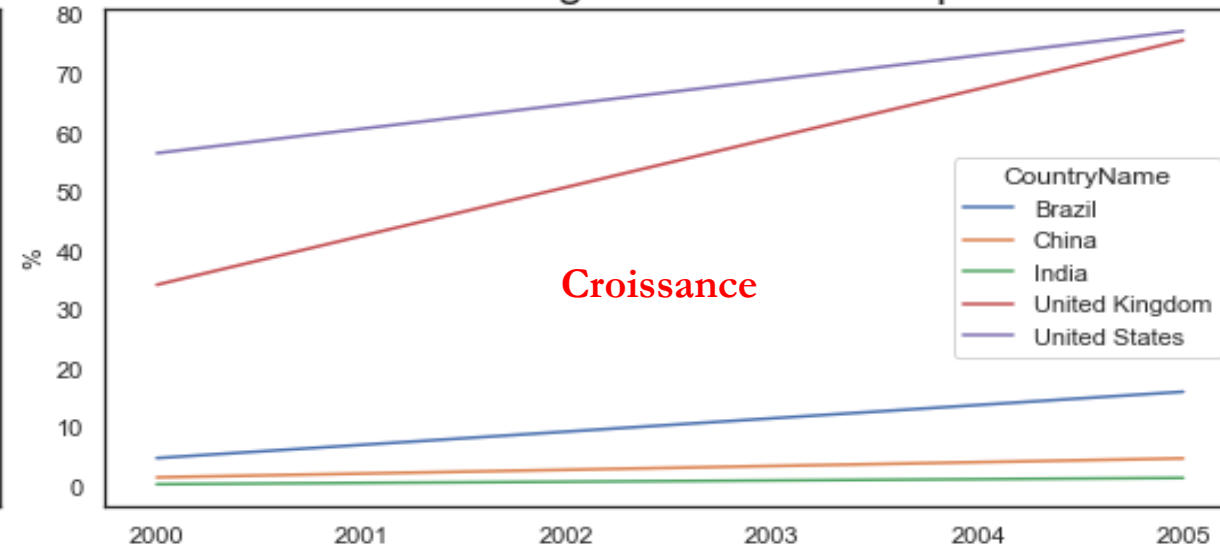


Evolution dans le temps

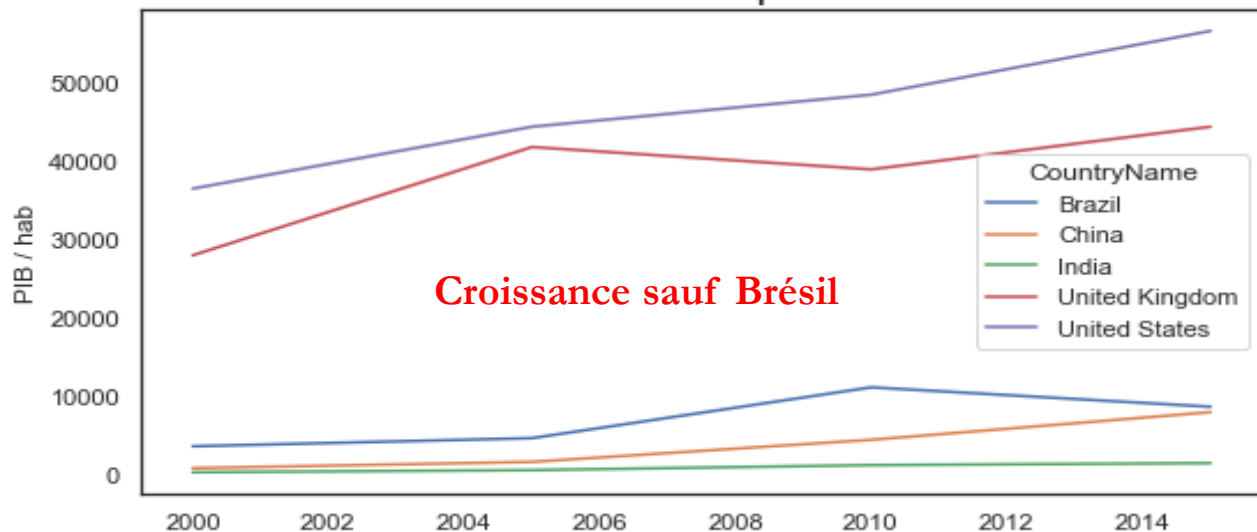
Evolution de l'accès à Internet



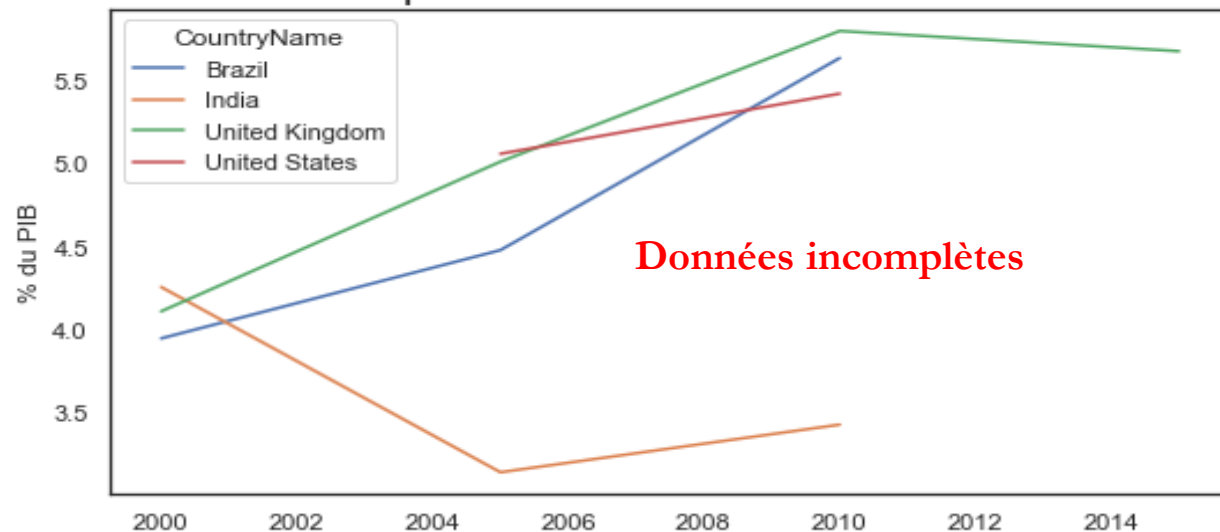
Evolution de l'usage d'un ordinateur personnel



Evolution du PIB par habitant

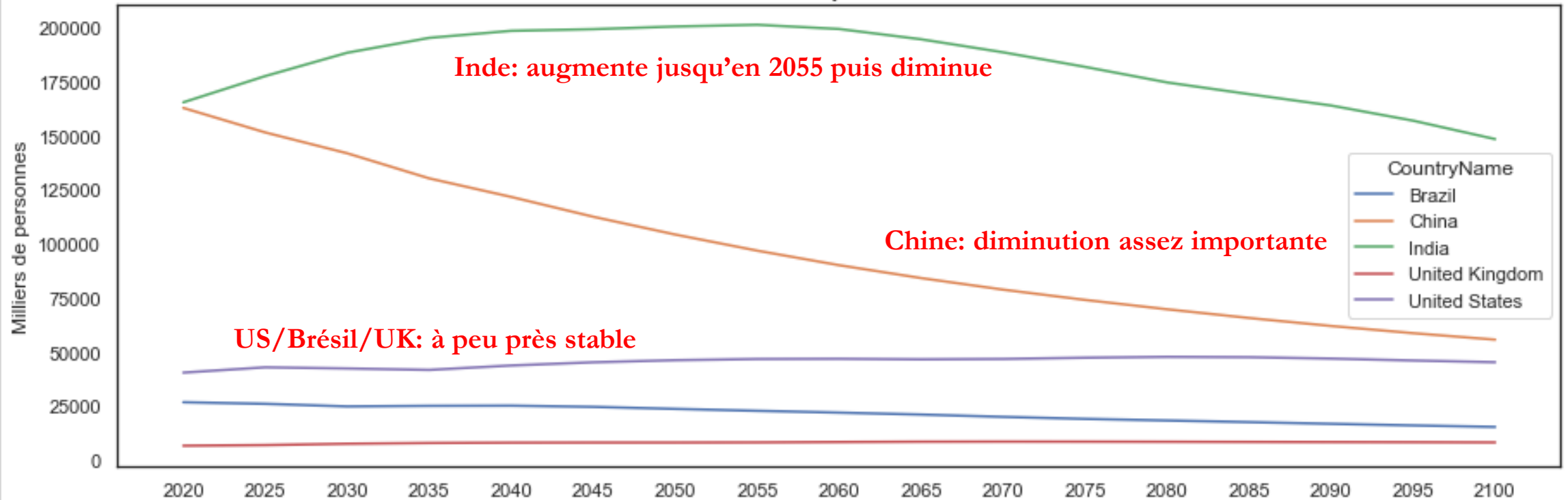


Evolution des dépenses du Gouvernement dans l'éducation



Focus indicateurs de projections

Wittgenstein Projection: Population 15-24 ans en milliers par plus haut niveau d'études:
secondaire et post secondaire





IV. PAYS CIBLES

Résultat scoring

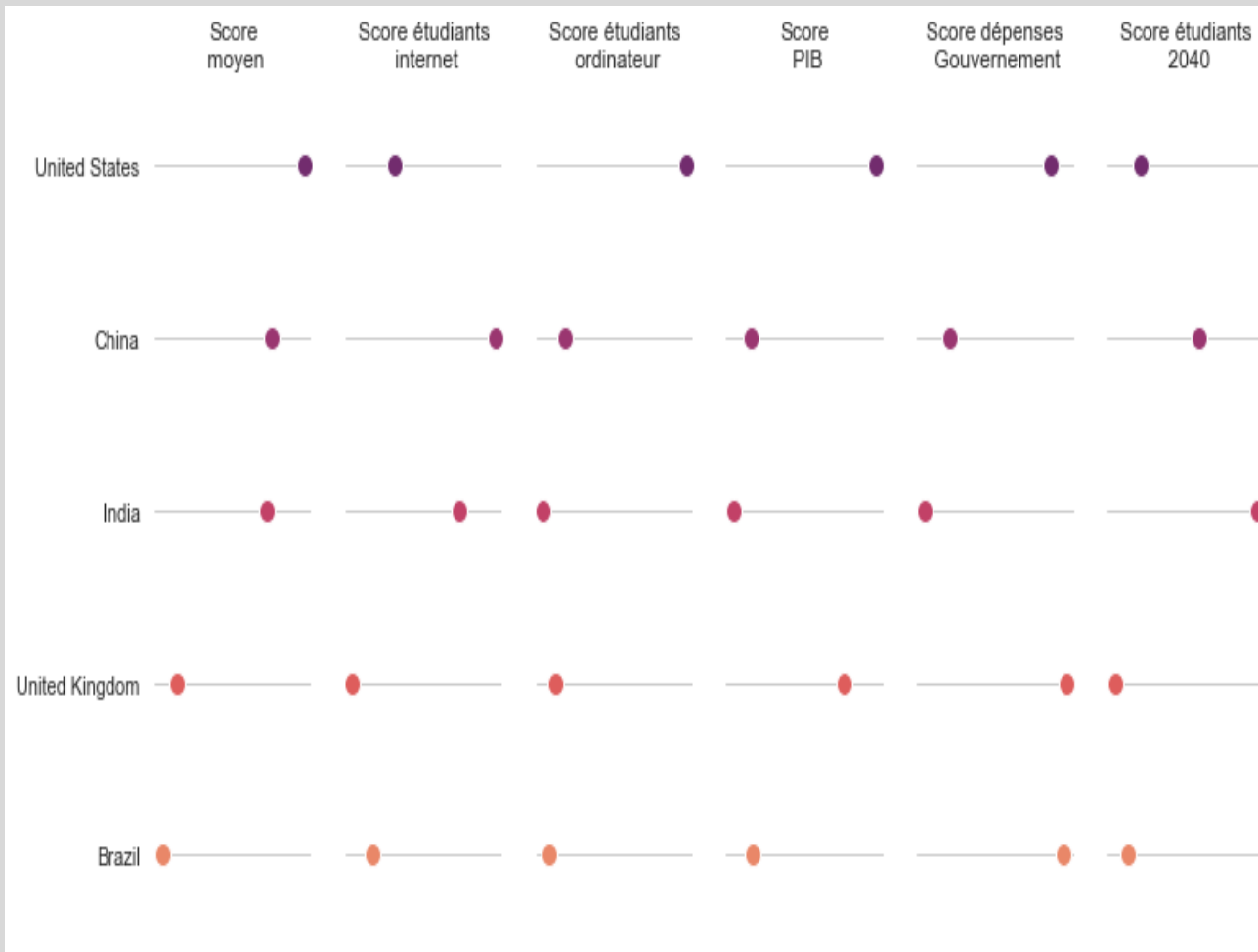
Application de la fonction:

- Population minimale de prospects ayant accès à internet: 4M
- SCORE_Propects_Internet = poids de 1
- SCORE_Propects_Ordinateurs= poids de 1
- SCORE_GDP_per_capita = poids de 0.5
- SCORE_Gov_expenditures_%_GDP= poids de 0.7
- SCORE_Pop_15-24_2040 = poids de 1

```
score_pays(classement_pays, 4000000, 1, 1, 0.5, 0.7, 1, 5)
```

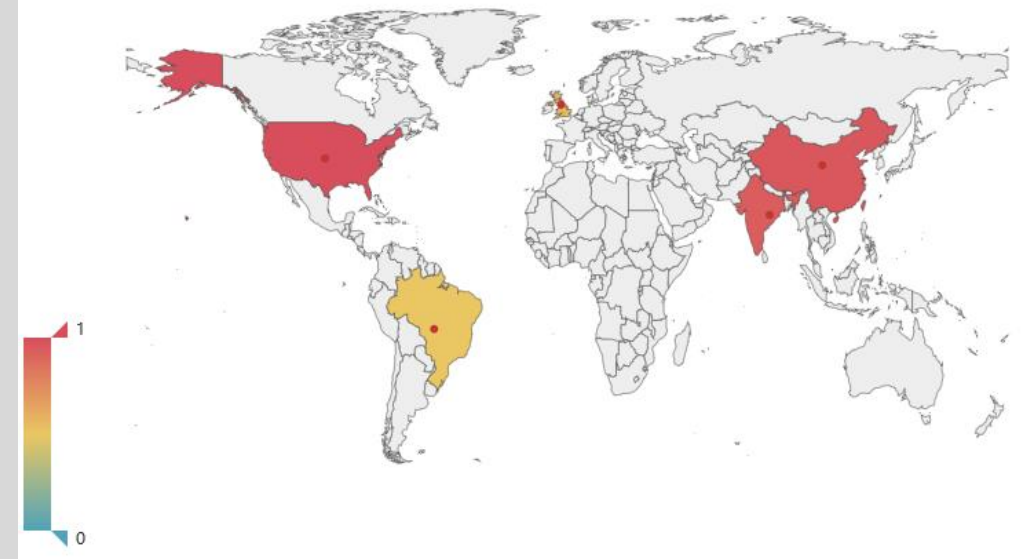
CountryName	SCORE_SYNT	SCORE_Propects_Internet	SCORE_Propects_Ordinateurs	SCORE_GDP_per_capita	SCORE_Gov_expenditures_%_GDP	SCORE_Pop_15-24_2040
United States	0.55	0.36	1.00	1.00	0.95	0.22
China	0.48	1.00	0.26	0.14	0.67	0.61
India	0.47	0.77	0.12	0.03	0.60	1.00
United Kingdom	0.28	0.09	0.20	0.78	0.99	0.04
Brazil	0.25	0.21	0.16	0.16	0.99	0.13

Résultat scoring



TOP 5 des pays retenus

en fonction du score moyen





V. PERTINENCE DU JEU DE DONNÉES

Points forts et limites



- Grand nombre de **pays**
- **Indicateurs pertinents**
- **Sources** et **modes de calculs** indiqués
- **Historique** des données depuis 1970 et **projections**



- Données réelles **anciennes** (2016 et seulement 2% de données disponibles)
- Certains indicateurs ne sont pas remplis par tous les pays ou sont complétés à des dates différentes (**manque d'harmonisation**)
- Certains indicateurs importants sont **obsolètes** (2005 pour l'accès à internet et à un ordinateur personnel)
- La maîtrise de la langue **anglaise** n'est pas reflétée dans les indicateurs
- Pas d'information sur la **concurrence** ni sur la **stratégie commerciale** d'academy



MERCI !