



**DATA SCIENTIST**



# Projet N° 2

## Analysez des données de systèmes éducatifs

**Elaborée par: Mariem Kchaou**  
**Encadrée par: Mr. Wilfried Josset**

**Avril 2022**

# Plan

- 01** Problématique
- 02** Description du jeu de données
- 03** Préparation du jeu de données
- 04** Sélection des indicateurs
- 05** Statistique des indicateurs choisis
- 06** Score
- 07** Prédictions
- 08** Conclusion

01

# Problématique

# 01

## Problématique

- Academy est une start-up de la EdTech
- Apprentissage en ligne: contenu de formation de niveau lycée et université
- Objectif d'expansion à l'international

A partir des données de la Banque mondiale, réaliser une pré-analyse exploratoire permettant de répondre aux interrogations suivantes :

- 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 ?



Pré-analyser des données de la banque mondiale



- Informations adéquates ? De qualité ? Consistantes
- Indicateurs statistiques pour arbitrer les pays
- Conclusion : Pays a fort potentiel ? Évolution ?

02

# Description du jeu de données

# 02

## Description du jeu de données

EdStatsCountry.csv

- Information géographique sur les pays par région
- 241 lignes, 32 variables, quelques valeurs manquantes, aucun doublon.

EdStatsSeries.csv

- Informations sur les indicateurs socio économique disponible
- 3665 lignes, 21 variables, plus de 80% de valeurs manquantes, aucun doublon

EdStatsFootNote.csv

- Information sur d'autres indicateurs par pays
- 241 lignes, 32 variables, pas de valeurs manquantes, aucun doublon

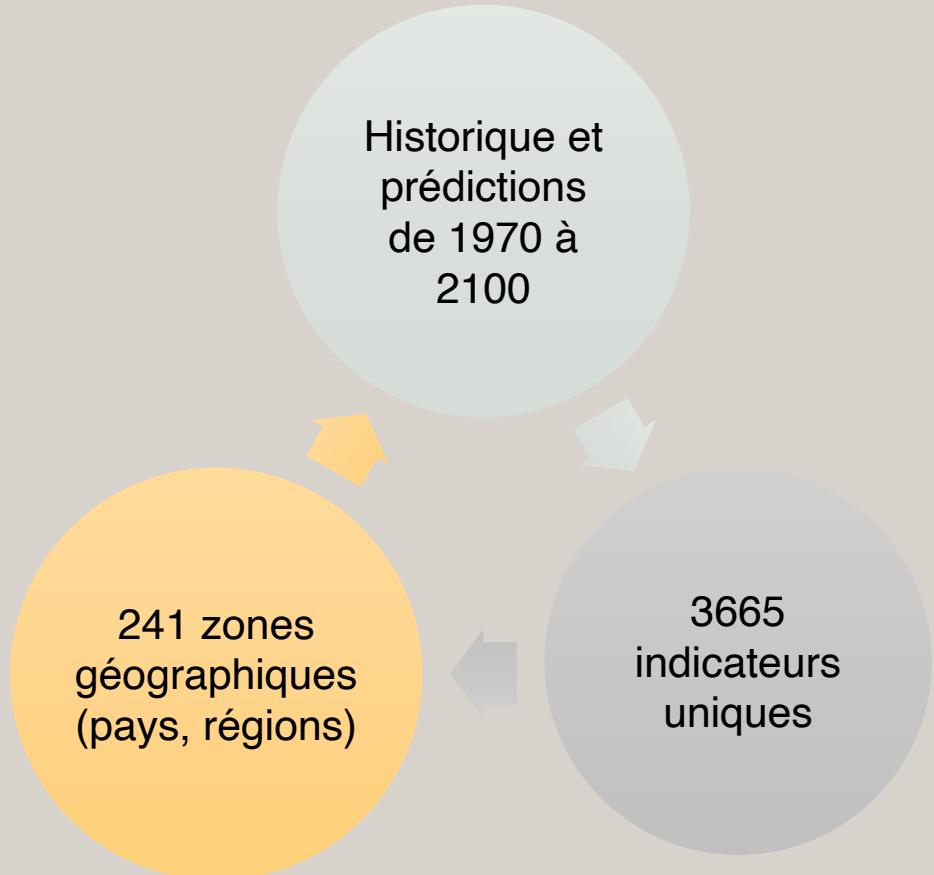
EdStatsCountry-Series.csv

- Informations sur les sources des indicateurs par pays
- 613 lignes, 4 variables, pas de valeurs manquantes, aucun doublon

EdStatsData.csv

- Evolution de nombreux indicateurs pour tout les pays
- 886930 lignes, 70 variables, entre 70% et 100% de valeurs manquantes, aucun doublon

## Le jeu de données « EdStatsData.csv »



- Le fichier « EdStatsData.csv » est le plus pertinent pour notre analyse. Les indicateurs par pays semblent pouvoir répondre à notre problématique
- Faire une sélection des données à garder, des indicateurs les plus renseignés, et les pays à fort potentiel.

Le jeu de données  
« EdStatsData.csv »

## Les outils

Environnement



Librairies de base



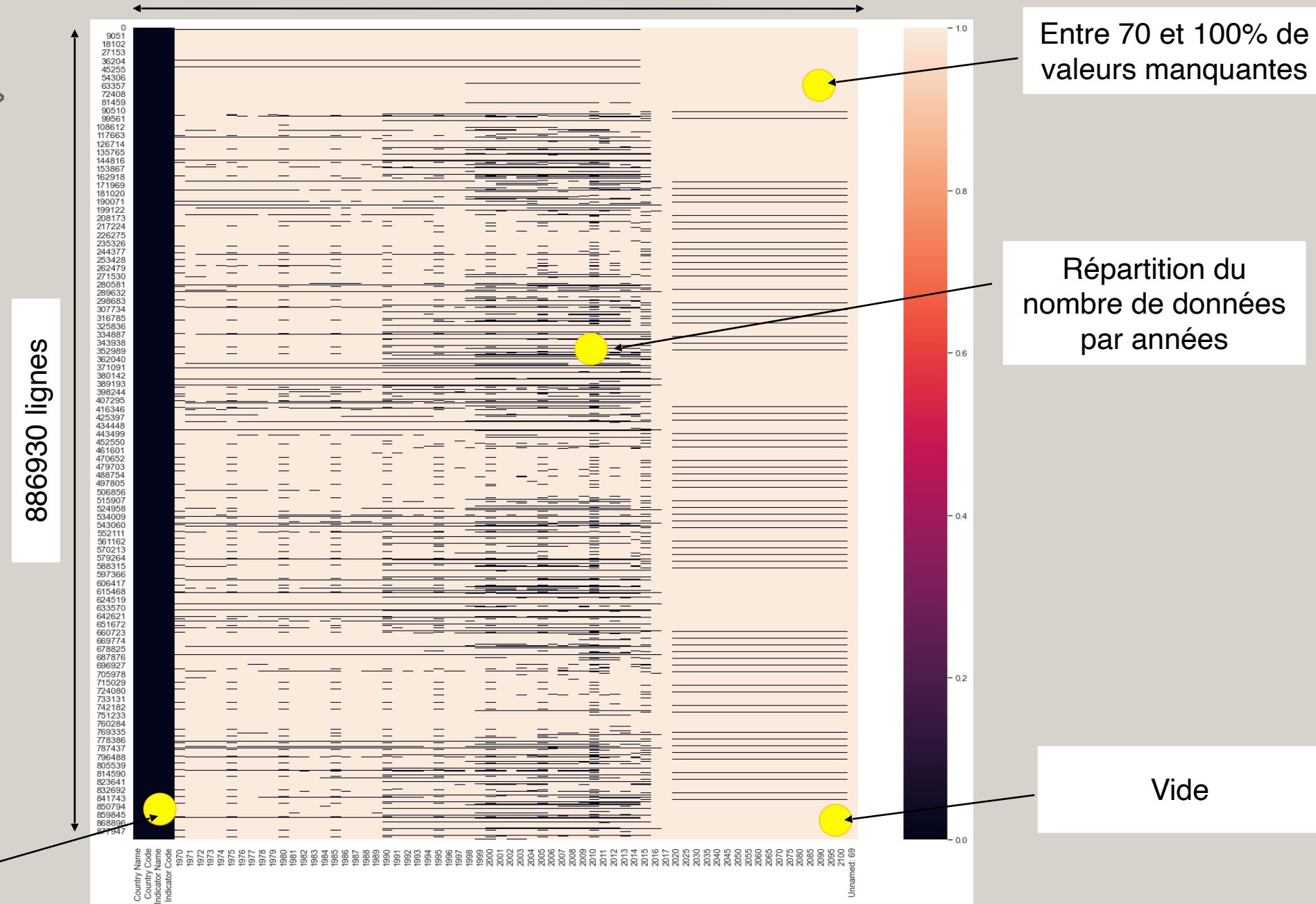
Visualisation



70 variables

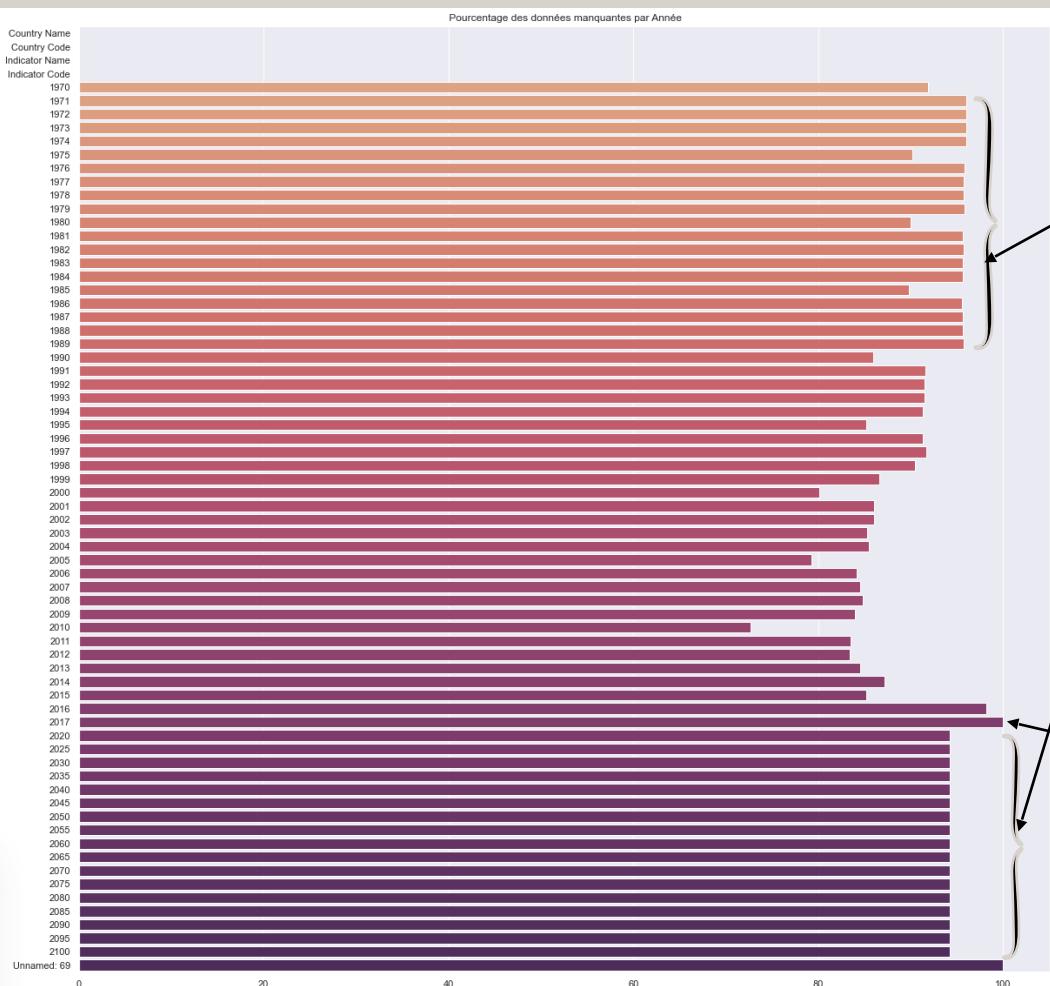
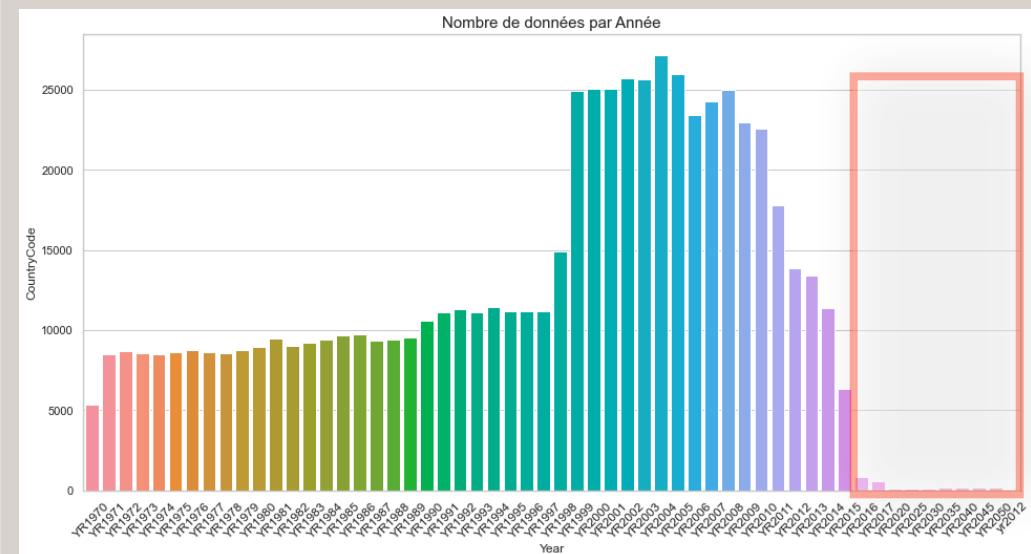
Le jeu de données  
« EdStatsData.csv »

Graphique indiquant  
les valeurs manquantes  
du notre dataframe



Le jeu de données  
« EdStatsData.csv »

## Pourcentage des données manquantes par années

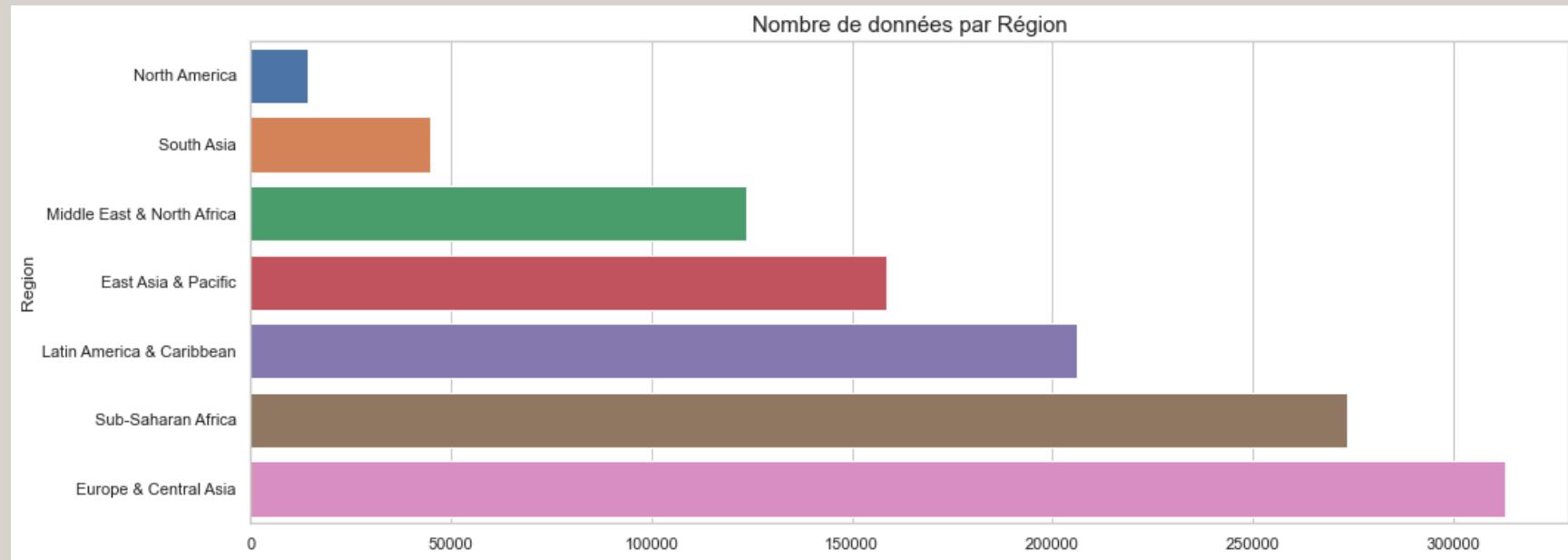


Nombre de  
données par  
années

Les années qui ont  
plusieurs de valeurs  
manquantes

Pas de données  
disponible

# Le jeu de données « EdStatsData.csv »



On voit qu'on a beaucoup plus de données pour l'Europe et l'Afrique sub saharienne que l'Amérique du nord ou l'Asie du Sud.

On peut expliquer ce résultat par le nombre de pays dans chaque région.

# 03

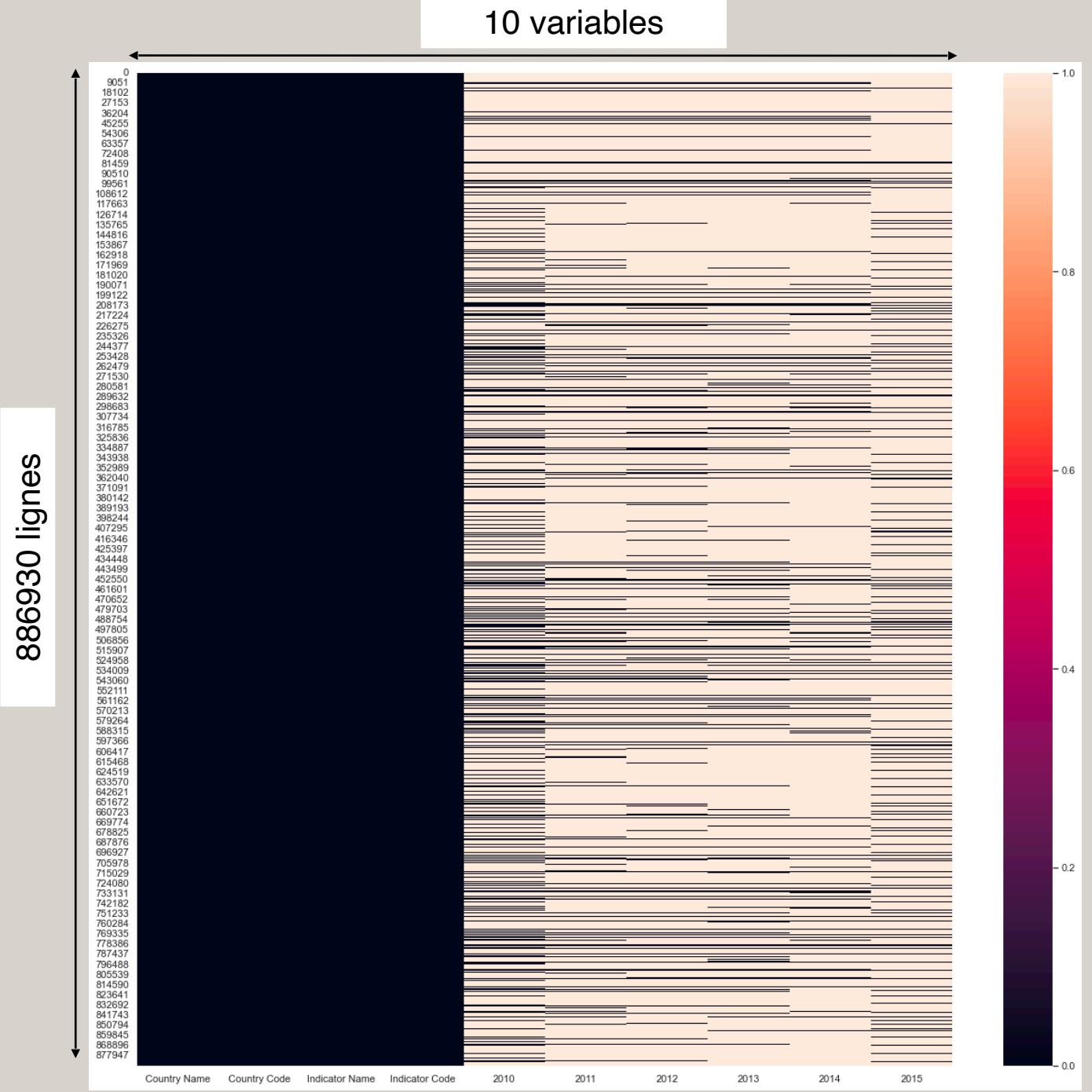
## Préparation du Jeu de données

# 03

## Préparation du jeu de données

Suppression des colonnes de '1970' à '2009'  
et de '2016' à '2100'.

On a maintenant une dataframe beaucoup  
plus renseigné.



04

# Sélection des indicateurs

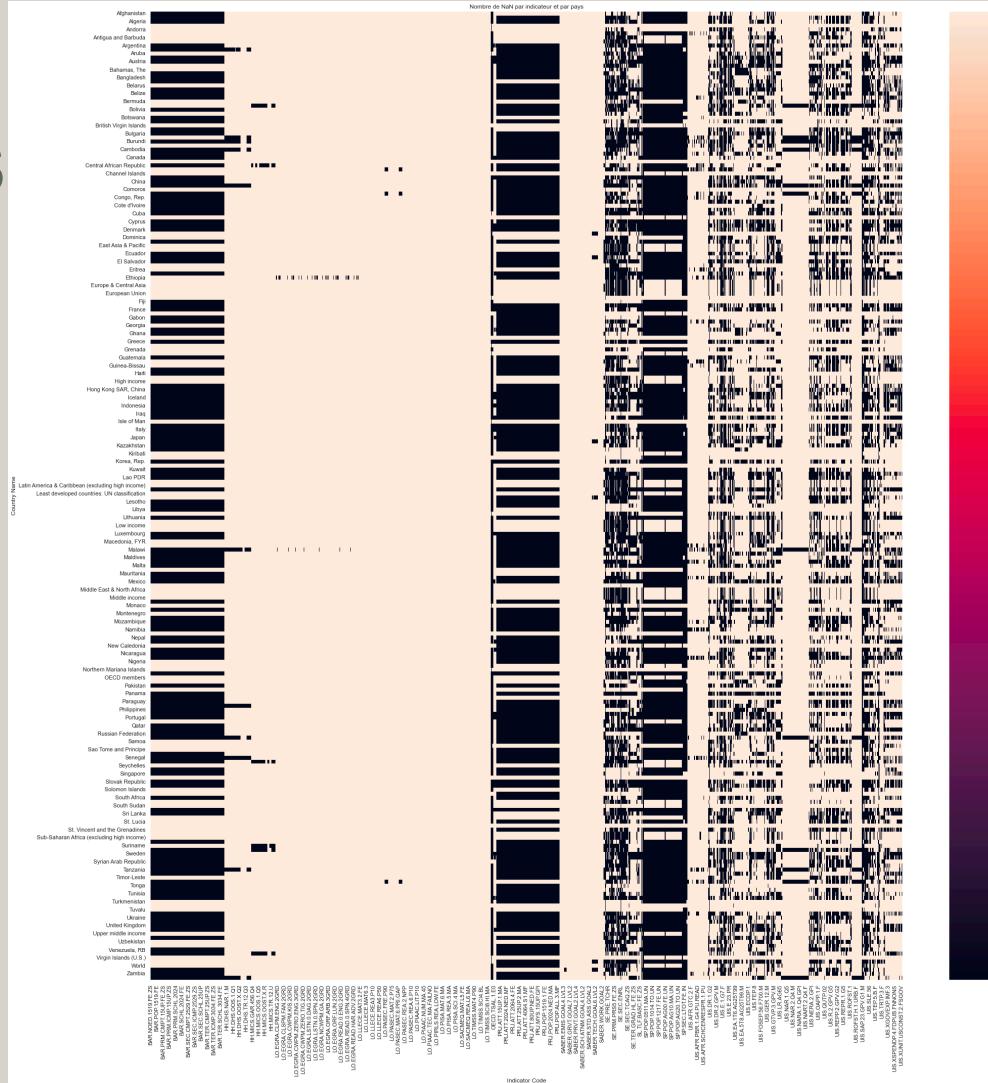
# 04

## Sélection des indicateurs

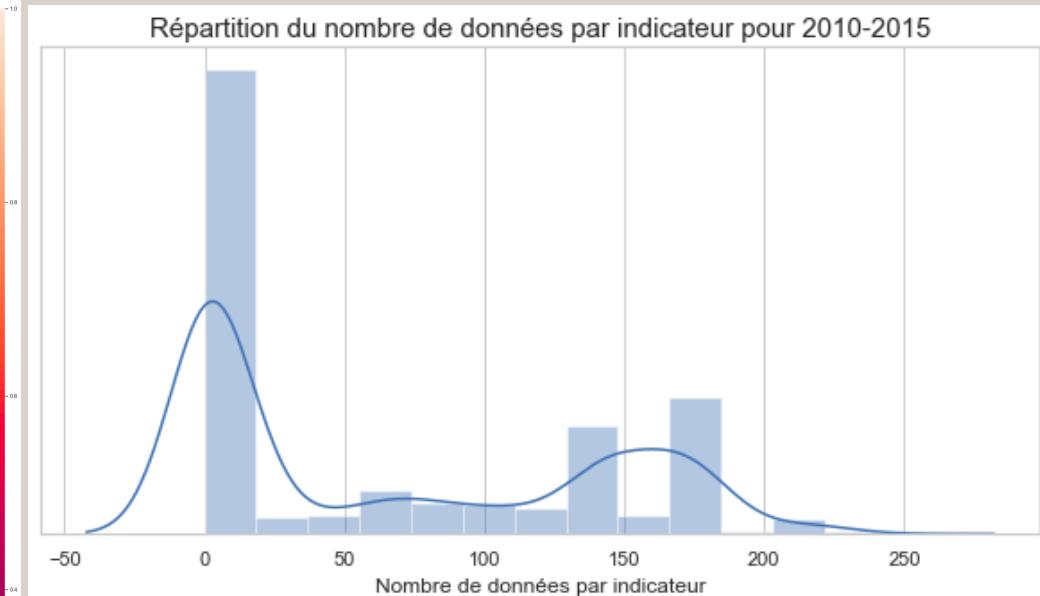
3665 indicateurs à analyser

### Les données manquantes

Données par indicateurs et par pays



Répartition du nombre de données par indicateur pour 2010-2015

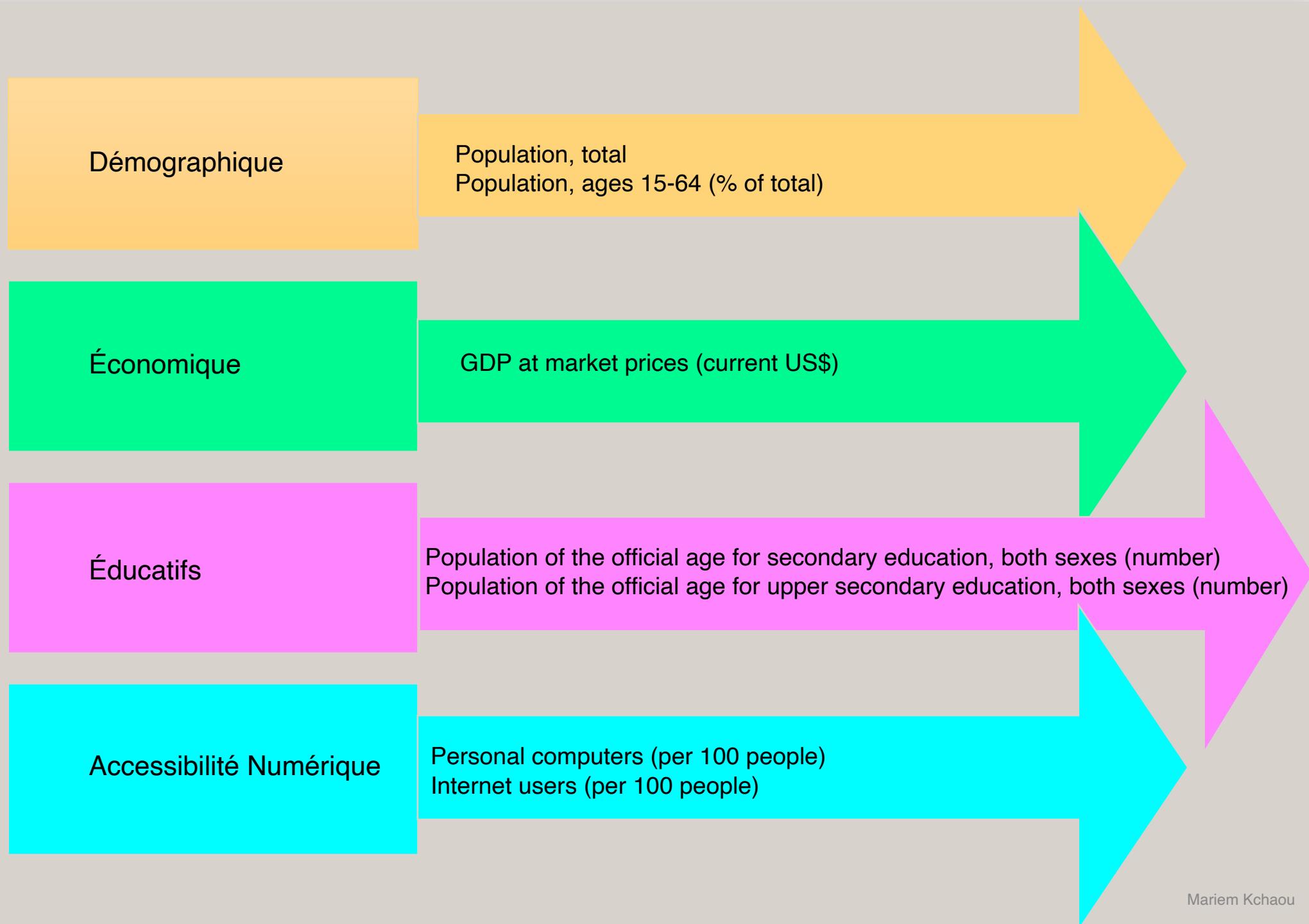


- Beaucoup de données manquantes
  - Identification des indicateurs avec le moins de données manquantes.
- On peut donc sélectionner les indicateurs les plus intéressants.



# 04

## Sélection des indicateurs



05

# Statistique des indicateurs choisis

# 05

## Statistique des indicateurs choisis

3665 indicateurs

7 indicateurs

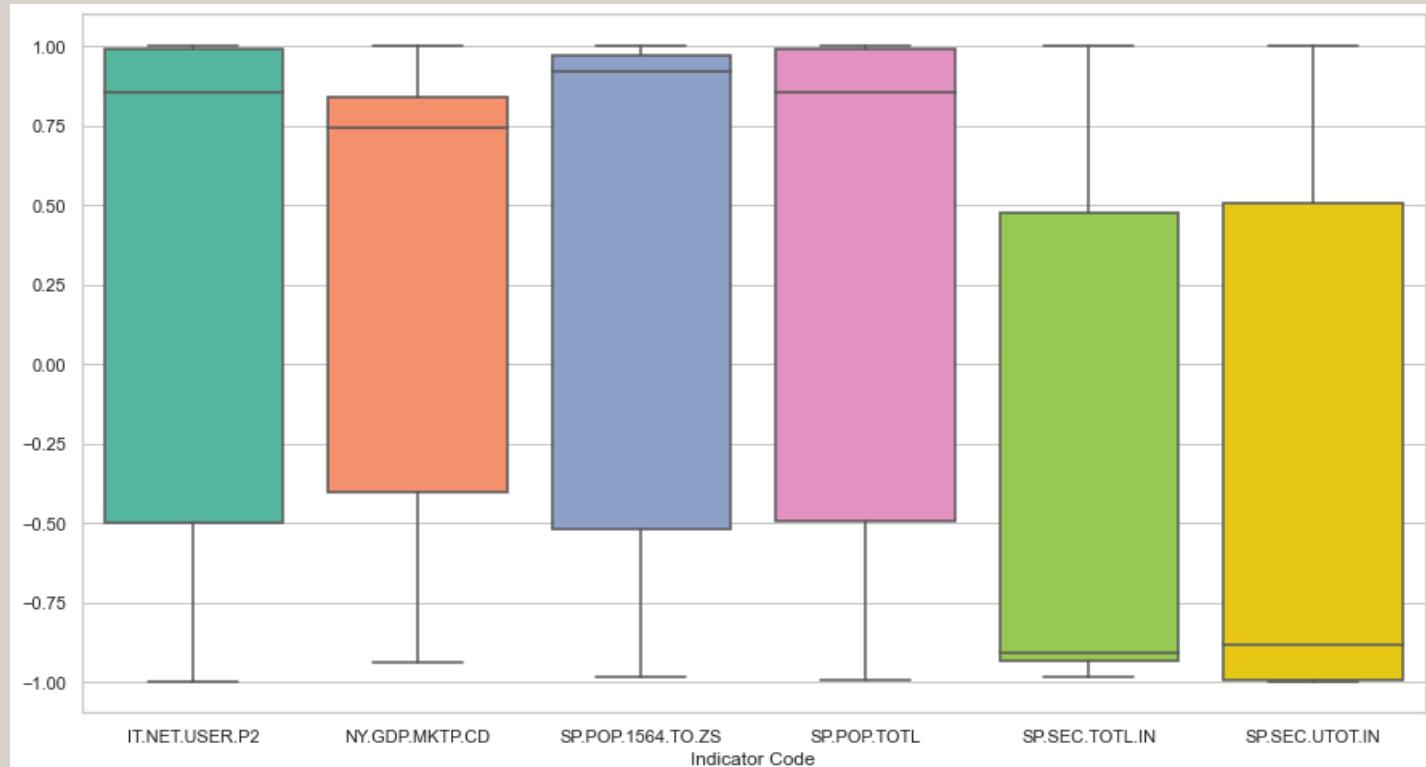
On constate que l'indicateur " Personal computers (per 100 people)" ne contient aucune données, tout pays confondus et pour toute la période de notre analyse.  
Nous pouvons donc l'écarter de notre analyse

6 indicateurs

### Statistique des indicateurs choisis

	2010	2011	2012	2013	2014	2015
count	1.294000e+03	1.294000e+03	1.294000e+03	1.294000e+03	1.294000e+03	1.294000e+03
mean	5.007969e+10	5.578694e+10	5.697554e+10	5.858738e+10	6.007138e+10	5.679440e+10
std	5.153429e+11	5.549701e+11	5.778956e+11	5.969789e+11	6.228380e+11	6.340660e+11
min	-2.270802e+00	-2.628656e+00	-2.147571e+00	-3.040564e+00	-3.107229e+00	-2.467847e+00
25%	3.737750e+01	4.053319e+01	4.695150e+01	4.992338e+01	5.377250e+01	5.760872e+01
50%	2.468580e+05	2.455230e+05	2.478855e+05	2.467785e+05	2.464625e+05	2.430710e+05
75%	2.469986e+06	2.467520e+06	2.482565e+06	2.478260e+06	2.469134e+06	2.485455e+06
max	1.496437e+13	1.551793e+13	1.615526e+13	1.669152e+13	1.739310e+13	1.812071e+13

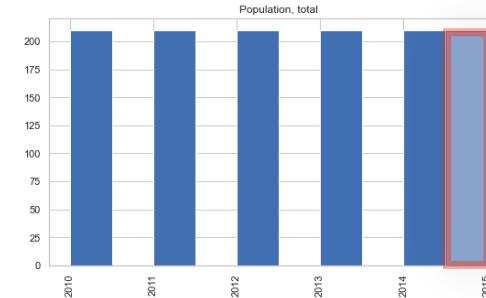
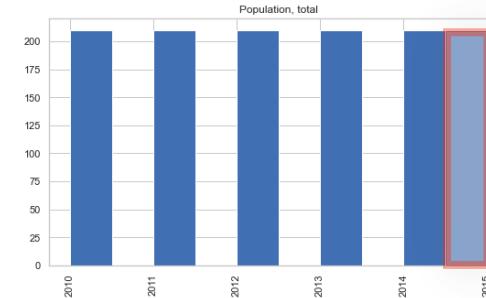
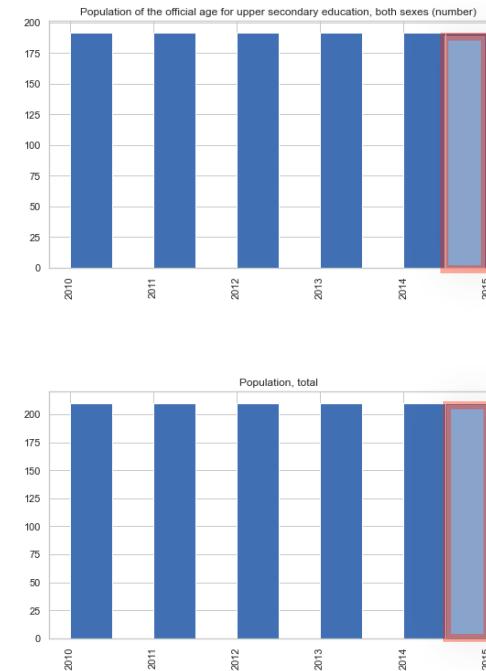
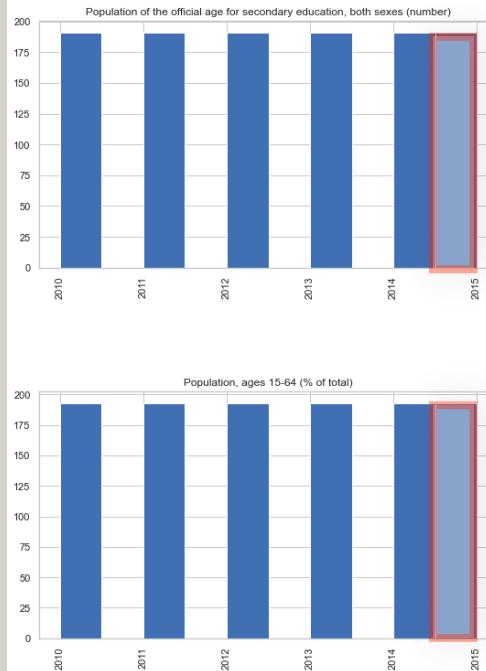
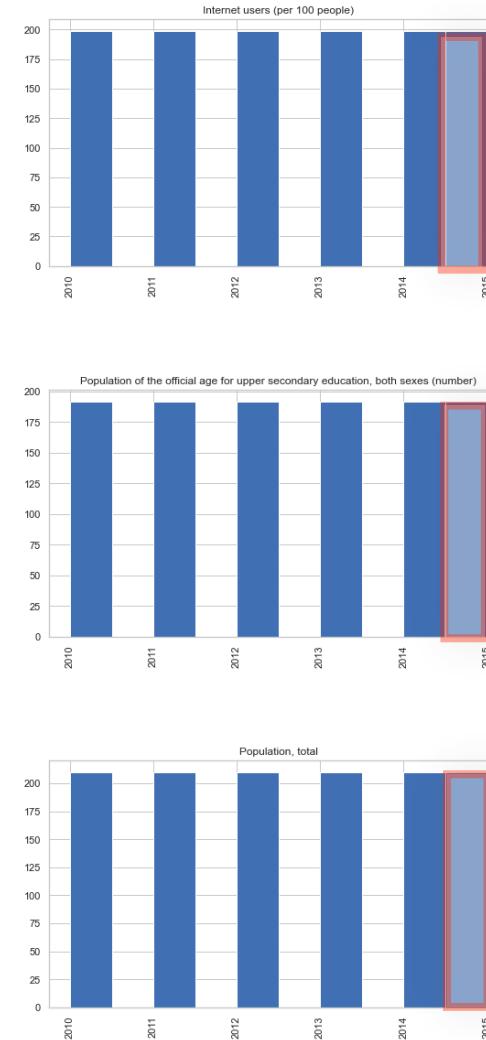
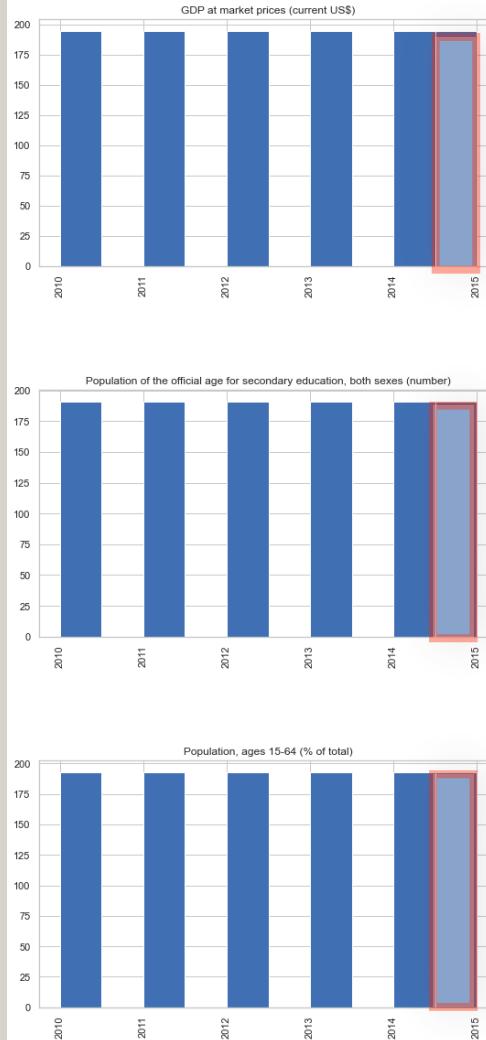
### Boxplot des données des indicateurs choisis



# 05

## Statistique des indicateurs choisis

### Histogramme des indicateurs choisis par année

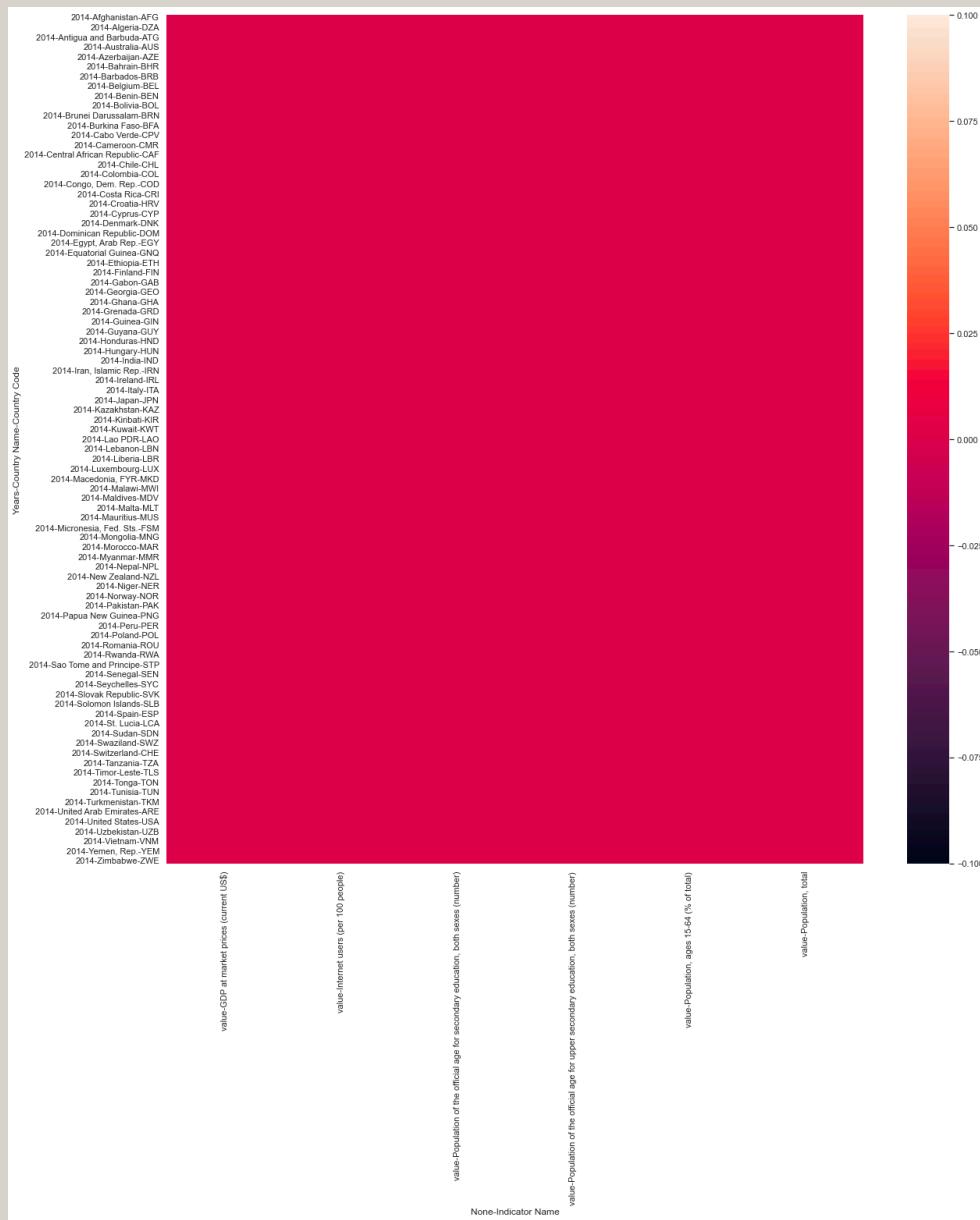


On a constaté que l'année 2015 est l'année la plus remplie.  
Donc on va terminé notre analyse avec l'année 2015

# 05

## Statistique des indicateurs choisis

### Graphique après la suppression des valeurs manquantes



Dernier nettoyage du notre datafarmame  
selon les indicateurs choisis

06

# Score

# 06

## Score

	Years	Country Name	Country Code	value		Total_Student	Total_Population
Indicator Name				GDP at market prices (current US\$)	Internet users (per 100 people)		
0	2015	Afghanistan	AFG	1.921556e+10	8.260000	7155249.0	3.373655e+07
1	2015	Albania	ALB	1.133526e+10	63.252933	485552.0	2.880772e+06
2	2015	Algeria	DZA	1.658743e+11	38.200000	5852743.0	3.987159e+07
3	2015	Angola	AGO	1.029622e+11	12.400000	5237369.0	2.785936e+07
4	2015	Antigua and Barbuda	ATG	1.364863e+09	70.000000	11007.0	9.999186e+04
...	...	...	...	...	...	...	...
168	2015	Vietnam	VNM	1.932411e+11	43.500000	13497705.0	9.171337e+07
169	2015	West Bank and Gaza	PSE	1.267300e+10	57.424192	1081304.0	4.422200e+06
170	2015	Yemen, Rep.	YEM	3.773392e+10	24.085409	5505009.0	2.691626e+07
171	2015	Zambia	ZMB	2.115439e+10	21.000000	2873711.0	1.610064e+07
172	2015	Zimbabwe	ZWE	1.630467e+10	22.742818	3389654.0	1.577751e+07

173 rows × 7 columns

An arrow points from the text "Somme de ces 2 indicateurs pour obtenir total des étudiants:" to the "Total\_Student" column header in the table.

An arrow points from the text "Somme de ces 2 indicateurs pour obtenir total du population:" to the "Total\_Population" column header in the table.

Somme de ces 2 indicateurs pour obtenir total des étudiants:  
 « Population of the official age for secondary education, both sexes (number) », et « Population of the official age for upper secondary education, both sexes (number) »

Somme de ces 2 indicateurs pour obtenir total du population:  
 « Population, total » et « Population, ages 15-64 (% of total) »

# 06

## Score

Years	Country Name	Country Code	Indicator Name				
			score_students	score_internet	score_population	score_finance	score_total
2015	Afghanistan	AFG	2.608283	8.411405	2.460331	0.106042	3.354621
	Albania	ALB	0.176997	64.412355	0.210088	0.062554	10.876111
	Algeria	DZA	2.133485	38.900204	2.907746	0.915385	8.224285
	Angola	AGO	1.909164	12.627291	2.031720	0.568202	3.593985
	Antigua and Barbuda	ATG	0.004012	71.283096	0.007292	0.007532	11.884426
	Argentina	ARG	2.286185	69.290289	3.166365	3.226757	13.605265
	Australia	AUS	0.838435	86.110509	1.734908	7.424559	15.673997
	Austria	AUT	0.382688	85.478760	0.629603	2.108449	14.729756
	Azerbaijan	AZE	0.430047	78.411405	0.703710	0.292893	13.438316

On calcule les scores des étudiants, d'internet, de population, d'économie et le score totale pour trouver les pays a fort potentiel

# 06

## Score

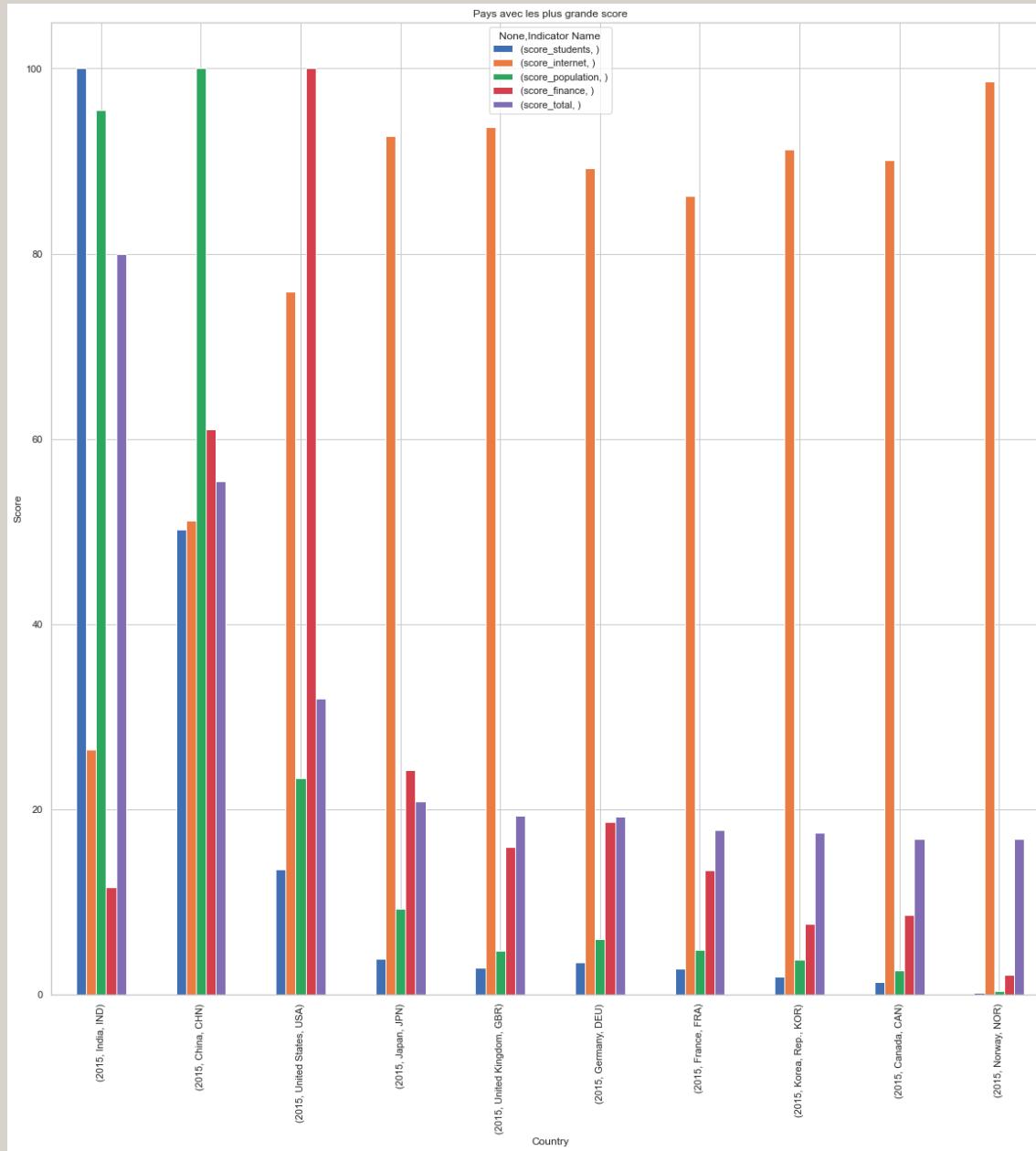
Years	Country Name	Country Code	Indicator Name				
			score_students	score_internet	score_population	score_finance	score_total
2015	India	IND	100.000000	26.476578	95.466371	11.533019	79.996046
	China	CHN	50.255729	51.221996	100.000000	61.060874	55.462558
	United States	USA	13.465058	75.920776	23.402274	100.000000	31.913691
	Japan	JPN	3.875252	92.727116	9.272112	24.188210	20.826381
	United Kingdom	GBR	2.920917	93.686660	4.749706	15.924153	19.284543
	Germany	DEU	3.418927	89.195316	5.957226	18.628466	19.193978
	France	FRA	2.812452	86.246945	4.858748	13.429725	17.773498
	Korea, Rep.	KOR	1.940386	91.291885	3.720411	7.630847	17.454843
	Canada	CAN	1.321094	90.091650	2.614364	8.569241	16.827971
	Norway	NOR	0.213324	98.584827	0.378398	2.133818	16.782372

Les 10 pays a fort potentiel: qui ont le score total la plus élevée .

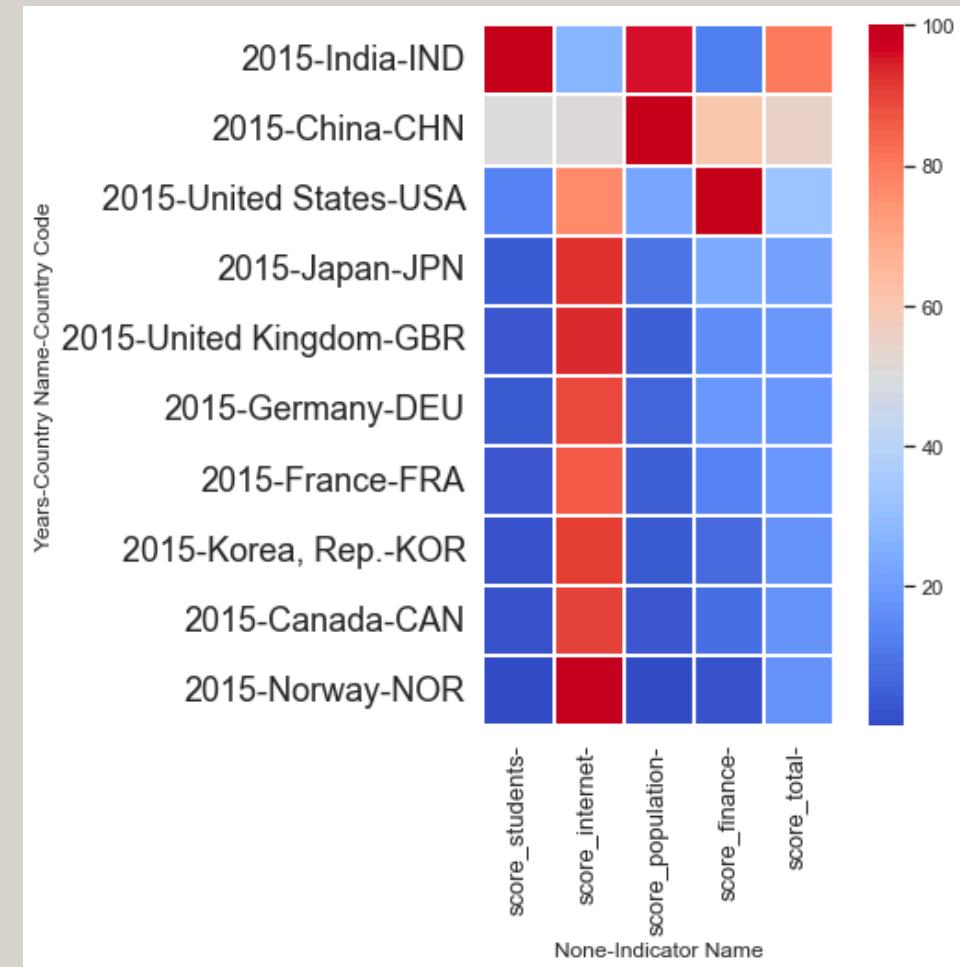
# 06

## Score

### Histogramme des pays a fort potentiel



### Heatmap des pays a fort potentiel



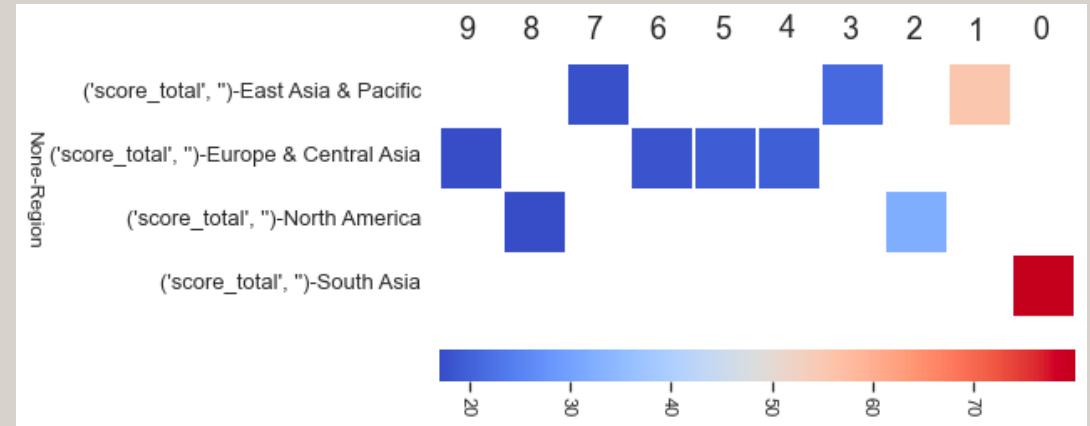
# 06

## Score

Nuage de points des régions a fort potentiel



Heatmap des régions a fort potentiel



(score_total, ) Region		
Country Code		
CHN	55.462558	East Asia & Pacific
JPN	20.826381	East Asia & Pacific
KOR	17.454843	East Asia & Pacific
GBR	19.284543	Europe & Central Asia
DEU	19.193978	Europe & Central Asia
FRA	17.773498	Europe & Central Asia
NOR	16.782372	Europe & Central Asia
USA	31.913691	North America
CAN	16.827971	North America
IND	79.996046	South Asia

Les 10 pays les plus attractifs font bien parties des 4 régions suivantes:

- Asie de l'Est et Pacifique
- Europe & Asie centrale
- Amérique du Nord
- Asie du Sud

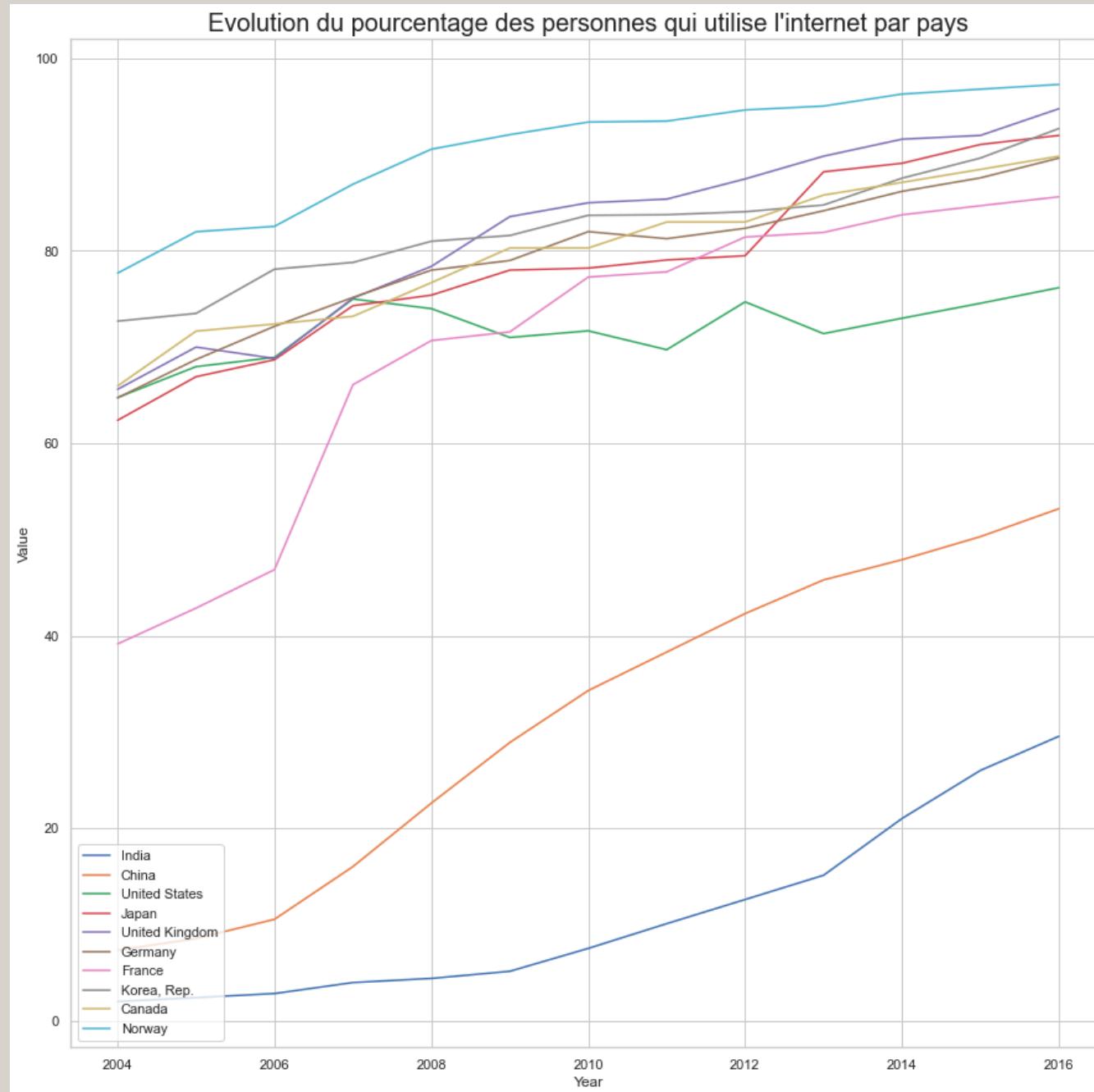
07

# Prédiction

# 07

## Prédiction

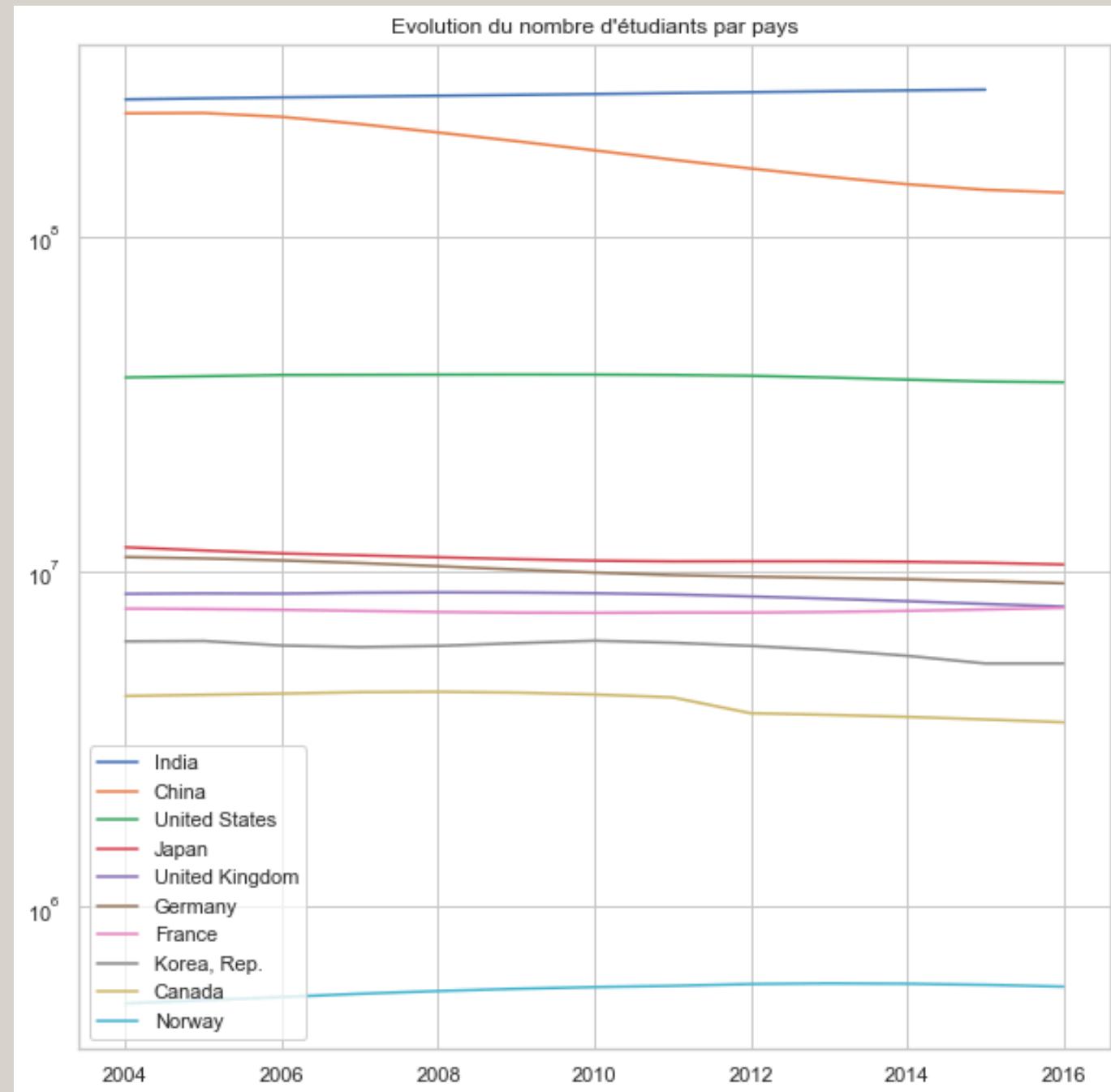
Diagramme du pourcentage des personnes qui utilise l'internet par pays des années « 2014-2016 »



# 07

## Prédiction

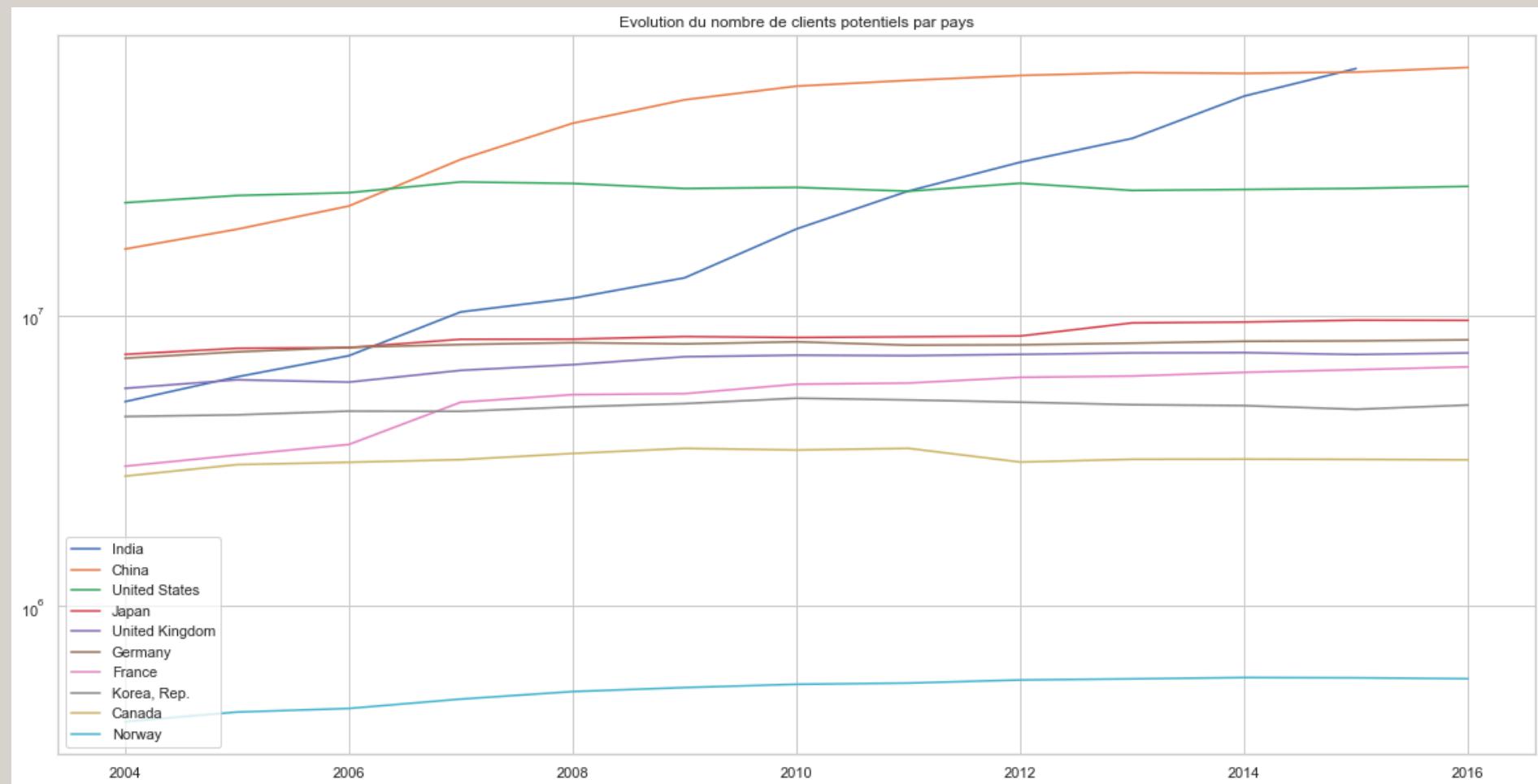
Diagramme d'évolution du nombre d'étudiants par pays des années « 2014-2016 »



# 07

## Prédiction

Diagramme d'évolution du nombre de clients potentiels par pays des années « 2014-2016 »



# Prédiction

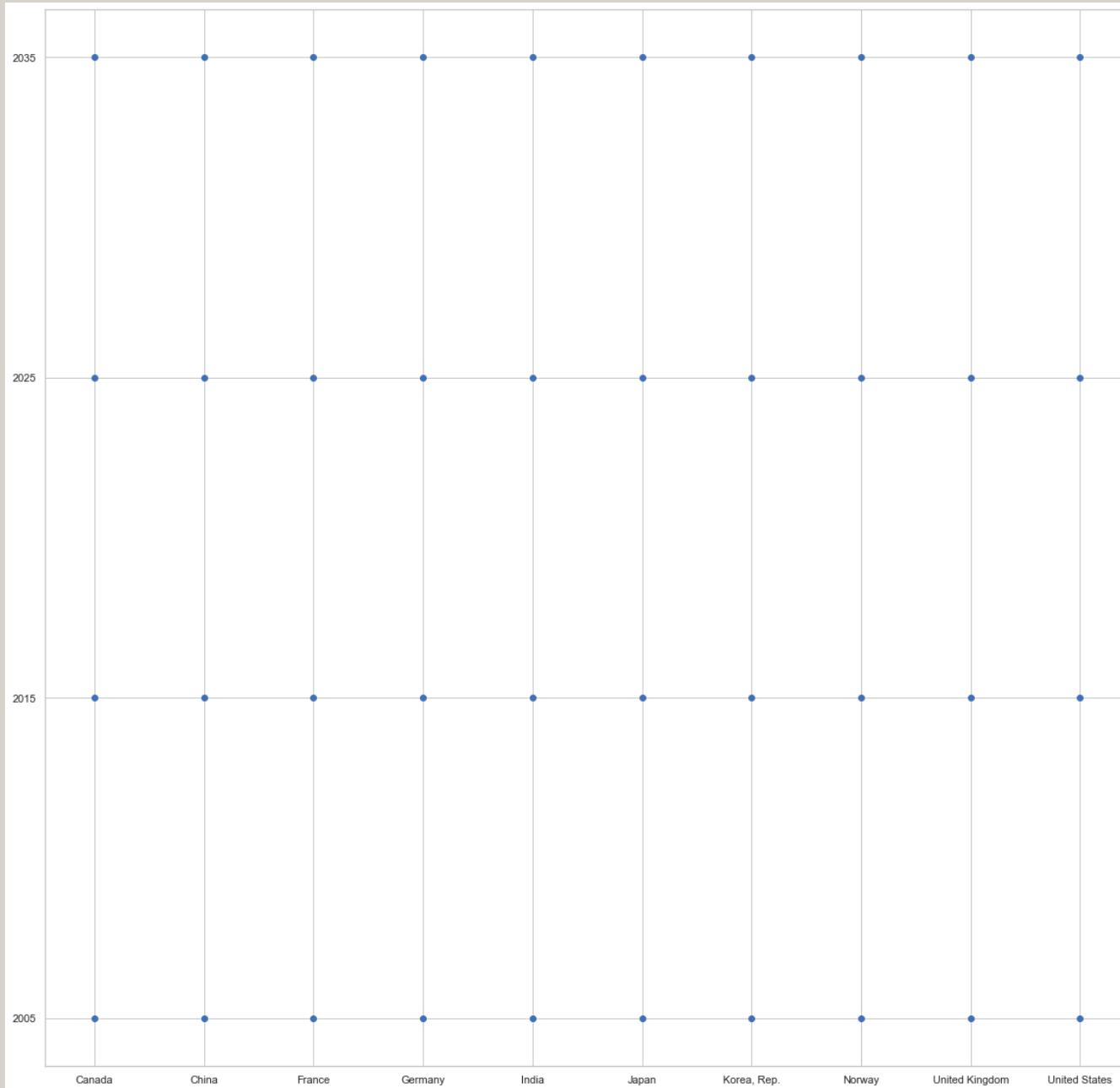
Les données de notre dataframe ne nous renseignent pas directement car on a pas des données pour le futur. Par contre on peut supposer que :

- Les pays qui ont une grande importance du nombre de clients vont continuer à voir plus d'importance dans les prochaines années.
- Les pays où l'internet moins implantés vont continuer à voir l'utilisation d'internet augmenter, et donc le nombre de clients.
- Les pays où la leur population sont importantes verront le nombre d'étudiants augmenter et donc le nombre de clients potentiels

# 07

## Prédiction

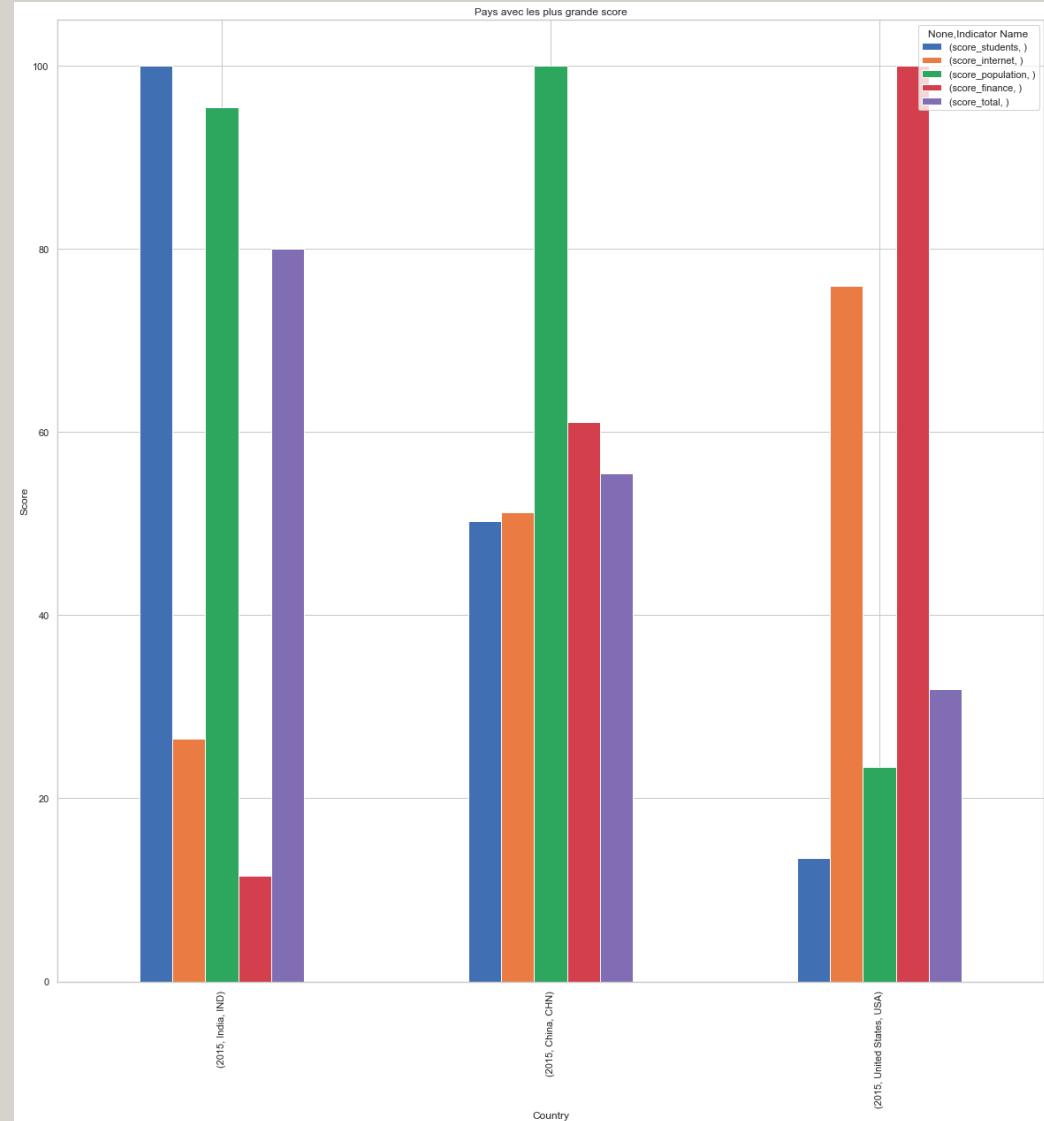
Nuage des points du prédition de nombre d'étudiants des années 2025 et 2035



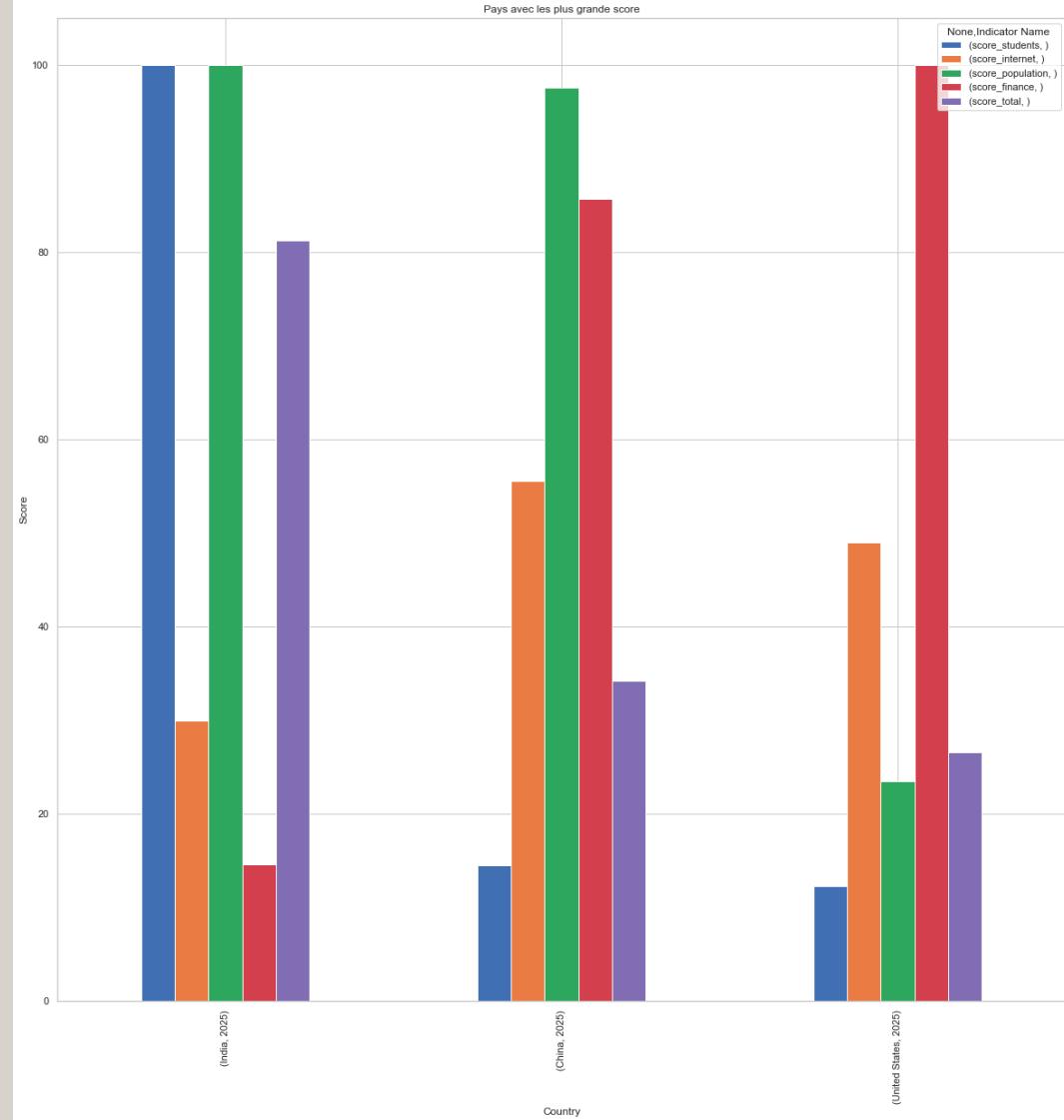
# 07

## Prédiction

Pays avec les plus grande score en 2015



Pays avec les plus grande score en 2025



On constate que les trois premières pays a fort potentiel en 2015 restent les même en 2025 qui sont l'Inde, la chine et l'Etats-Unis

08

# Conclusion

# 08

## Conclusion

### Les pays

Inde



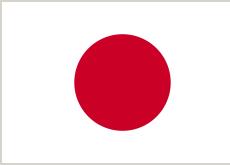
Chine



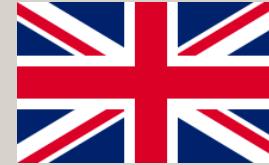
Etats-Unis



Japon



Royaume-Uni



Allemagne



France



Corée, Rép.



Canada



Norvège



### Les régions

Asie de l'Est et Pacifique

Europe & Asie centrale

Amérique du Nord

Asie du Sud

### Pour aller plus loin

- Jeu de données plus récent, plus de données éducatives.
- Forte impression des pays d'Asie de l'Est et du Pacifique.
- Aucune information sur la société Academy pour guider l'étude proximité géographique, concurrence, langue....

# 08

## Conclusion

### Pour aller plus loin

- Jeu de données plus récent, plus de données éducatives (les moyennes des étudiants, les notes...).
- Forte impression des pays d'Asie de l'Est et du Pacifique.
- Aucune information sur la société Academy pour guider l'étude proximité géographique, concurrence, langue....

Merci pour votre  
attention!

