P2 ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

Formation Data Scientist
OpenClassrooms

CONTEXTE

PROBLÉMATIQUE

Les données sur l'éducation de la banque mondiale permettent elles d'informer le projet d'expansion à l'international ?

OBJECTIFS

Décrire les informations contenues dans le jeu de données.

Sélectionner les informations qui semblent pertinentes pour répondre à la problématique.

Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde.

Valider la qualité de ce jeu de données.

DESCRIPTION DU JEU DE DONNÉES

- Sources : https://datacatalog.worldbank.org/dataset/education-statistics

- La base de données (EdStats) est composées de 5 tables :
  - EdStatsCountry.csv
  - EdStatsSeries.csv
  - EdStatsCountry-Series.csv
  - EdStatsFootNote.csv
  - EdStatsData.csv

DESCRIPTION DU JEU DE DONNÉES

EdStatsData.csv

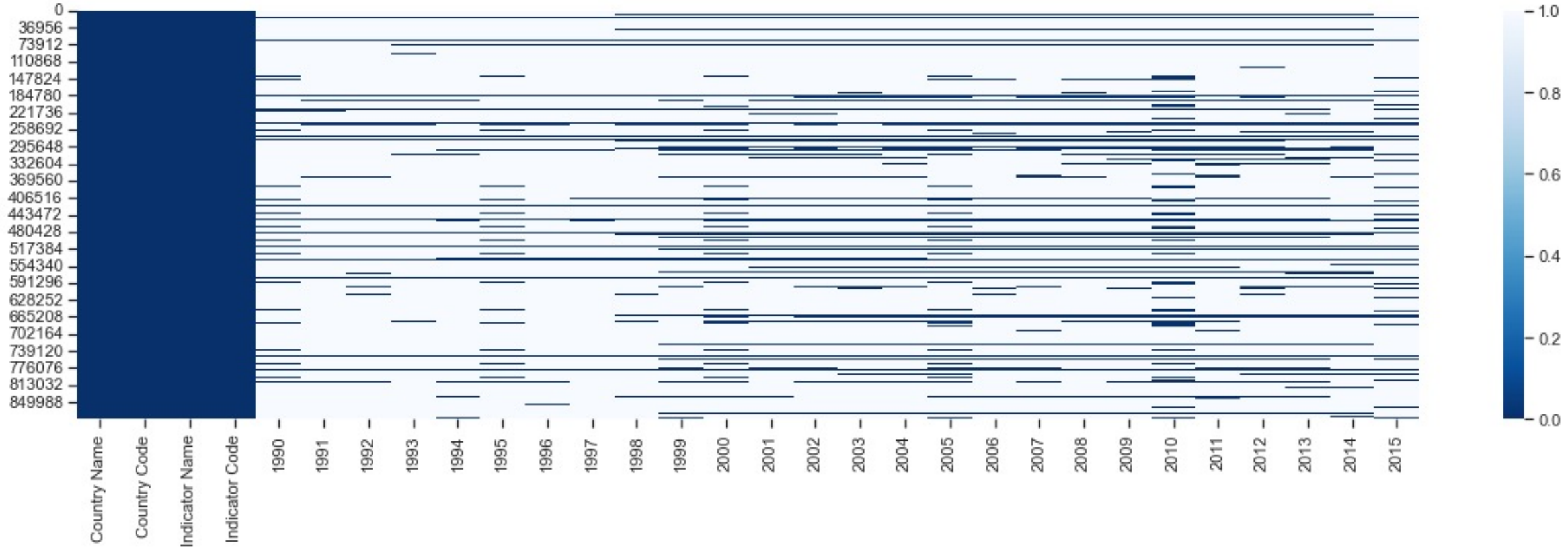
- 886 930 lignes / 70 colonnes ~ 62 millions de données

	Country Name	Country Code	Indicator Name	Indicator Code	1970	...	2085	2090	2095	2100	Unnamed: 69
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	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
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	...	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	...	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	...	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
886925	Zimbabwe	ZWE	Youth illiterate population, 15-24 years, male...	UIS.LP.AG15T24.M	NaN	...	NaN	NaN	NaN	NaN	NaN
886926	Zimbabwe	ZWE	Youth literacy rate, population 15-24 years, b...	SE.ADT.1524.LT.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
886927	Zimbabwe	ZWE	Youth literacy rate, population 15-24 years, f...	SE.ADT.1524.LT.FE.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
886928	Zimbabwe	ZWE	Youth literacy rate, population 15-24 years, g...	SE.ADT.1524.LT.FM.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
886929	Zimbabwe	ZWE	Youth literacy rate, population 15-24 years, m...	SE.ADT.1524.LT.MA.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN

886930 rows × 70 columns

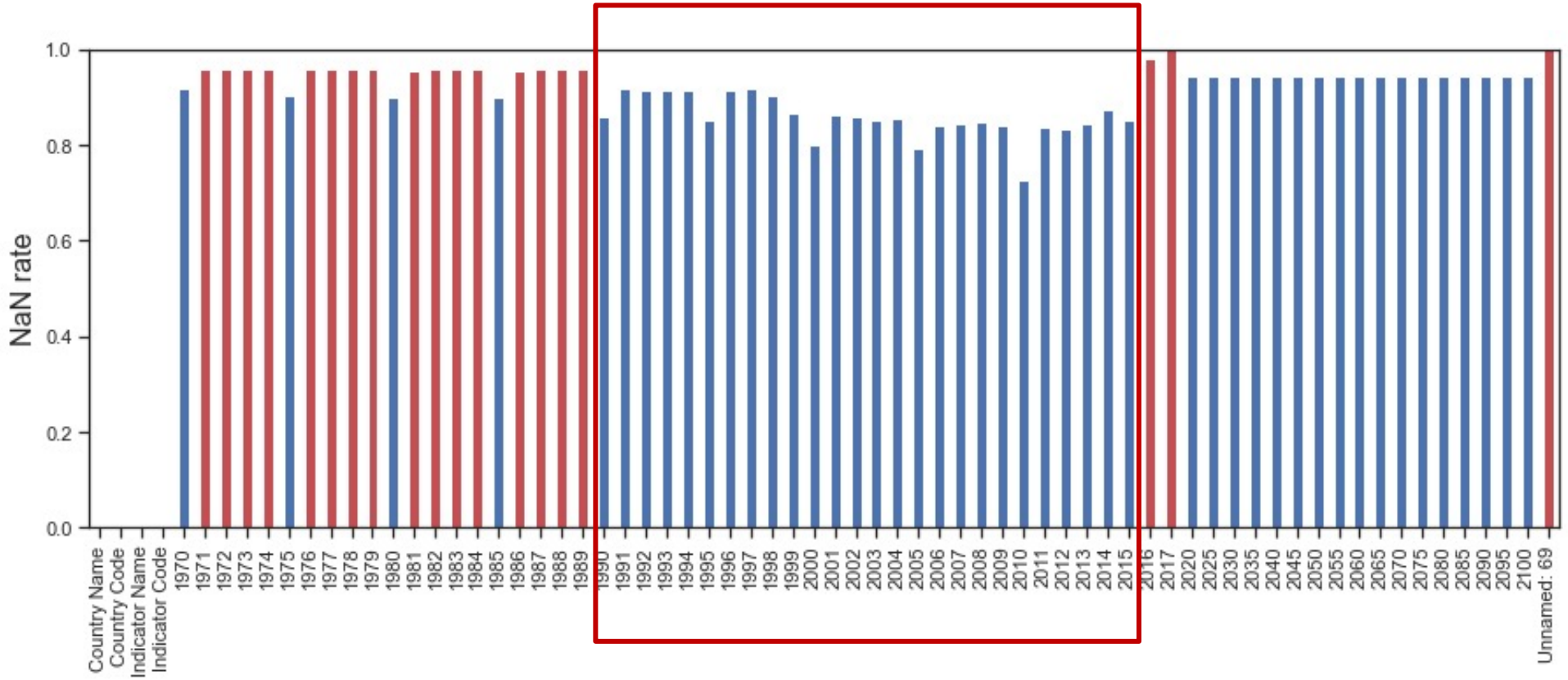
ANALYSE DES VALEURS MANQUANTES

- 86 % de valeurs manquantes



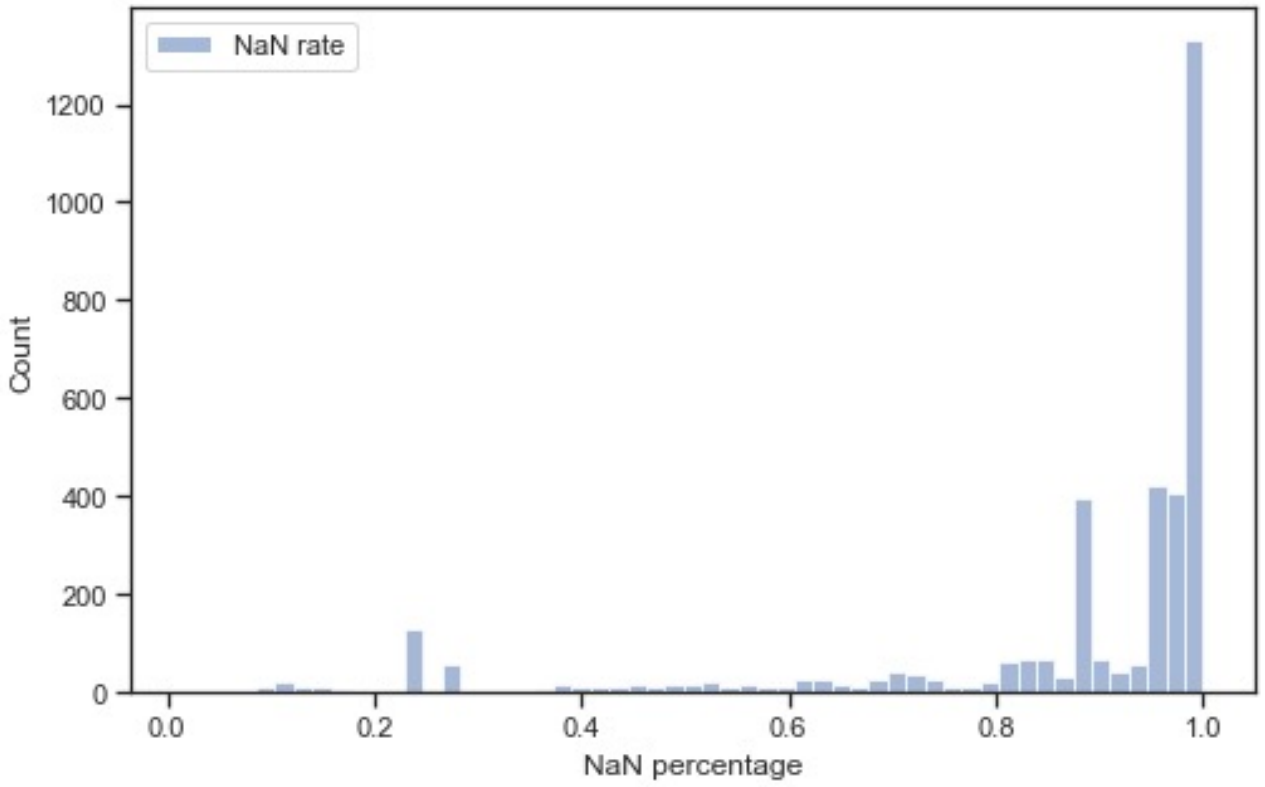
VALEURS MANQUANTES PAR COLONNE

- Le pourcentage de valeur manquante médian par colonne est 94 %. *(colonne avec un taux supérieur en rouge sur le graphe)*



VALEURS MANQUANTES PAR INDICATEUR

- Le pourcentage de valeur manquante médian par indicateur est 95 %.



Suppression des indicateurs avec un taux de remplissage < 50 % puis sélection manuelle

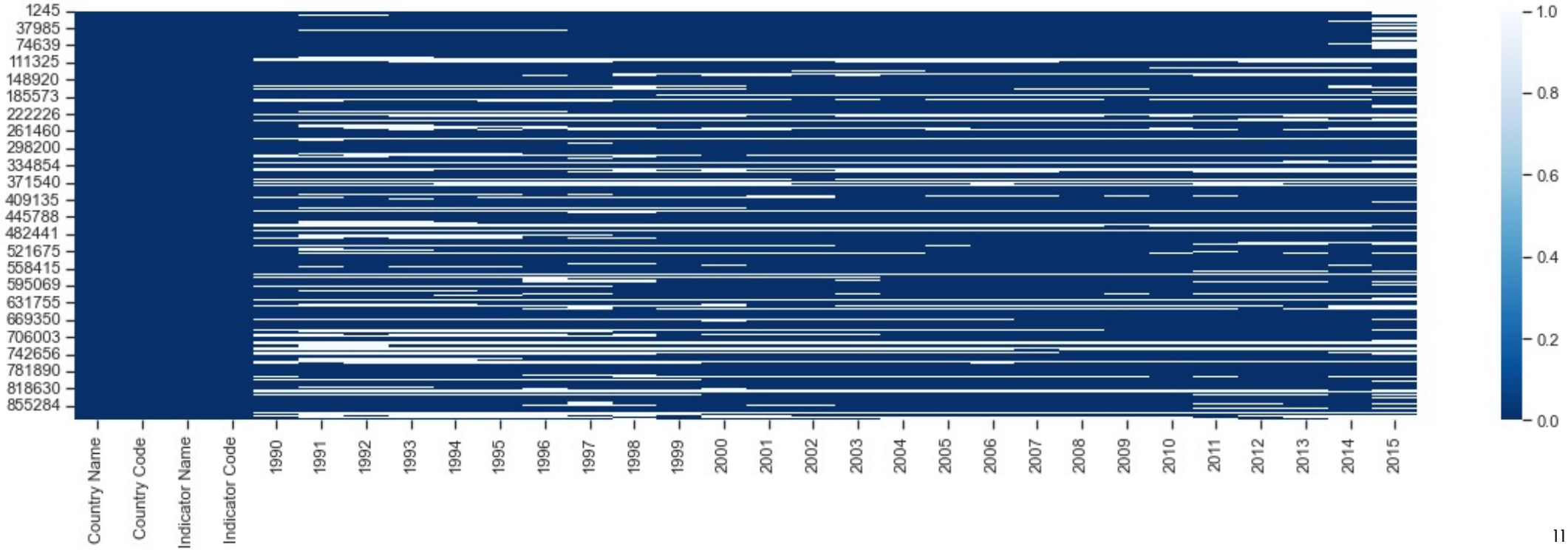
SÉLECTION D'INDICATEURS PERTINENTS

- **Economique** (valeur du marché)
- **Démographique** (taille du marché)
- **Infrastructure** (faisabilité)
- Efficacité système éducatif

Indicator Code	Indicator Name	NaN %	Importance
NY.GDP.PCAP.CD	GDP per capita (current US$)	9%	+
SP.SEC.TOTL.IN	Population of the official age for secondary education, both sexes (number)	11%	+
SP.PRM.TOTL.IN	Population of the official age for primary education, both sexes (number)	11%	+
SP.TER.TOTL.IN	Population of the official age for tertiary education, both sexes (number)	13%	+
IT.NET.USER.P2	Internet users (per 100 people)	19%	+
SE.SEC.ENRR	Gross enrolment ratio, secondary, both sexes (%)	35%	(+)
SE.TER.ENRR	Gross enrolment ratio, tertiary, both sexes (%)	41%	(+)

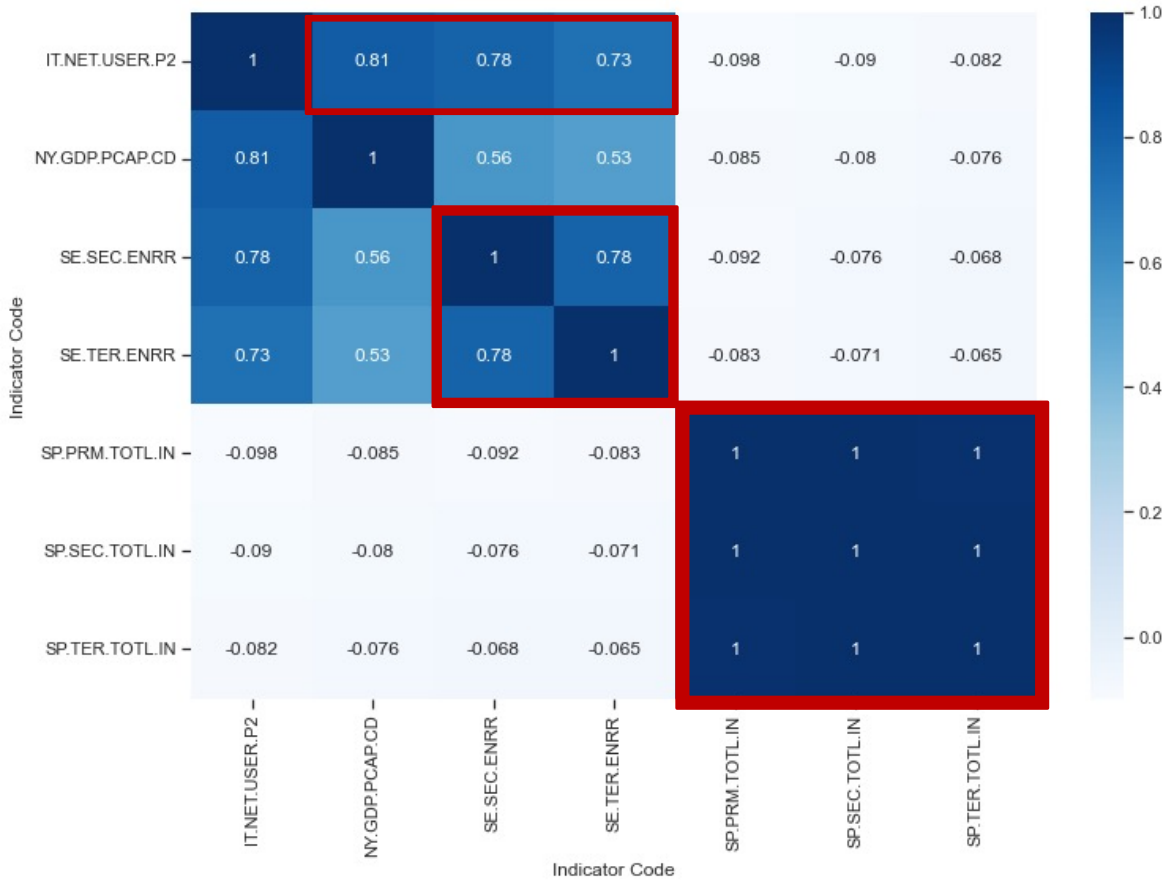
VALEURS MANQUANTES DU NOUVEAU JEU DE DONNÉES

- 1694 lignes et 30 colonnes ~ 50 milles données
- Plus que 17 % de NaN



ANALYSE DES INDICATEURS SÉLECTIONNÉS

Indicator Code	Indicator Name
NY.GDP.PCAP.CD	GDP per capita (current US$)
SP.SEC.TOTL.IN	Population of the official age for secondary education, both sexes (number)
SP.PRM.TOTL.IN	Population of the official age for primary education, both sexes (number)
SP.TER.TOTL.IN	Population of the official age for tertiary education, both sexes (number)
IT.NET.USER.P2	Internet users (per 100 people)
SE.SEC.ENRR	Gross enrolment ratio, secondary, both sexes (%)
SE.TER.ENRR	Gross enrolment ratio, tertiary, both sexes (%)



Création de nouveaux indicateurs en groupant

CRÉATION DE NOUVEAU INDICATEUR

Indicator Code	Indicator Name
NY.GDP.PCAP.CD	GDP per capita (current US$)
SP.SEC.TOTL.IN	Population of the official age for secondary education, both sexes (number)
SP.PRM.TOTL.IN	Population of the official age for primary education, both sexes (number)
SP.TER.TOTL.IN	Population of the official age for tertiary education, both sexes (number)
IT.NET.USER.P2	Internet users (per 100 people)
SE.SEC.ENRR	Gross enrolment ratio, secondary, both sexes (%)
SE.TER.ENRR	Gross enrolment ratio, tertiary, both sexes (%)

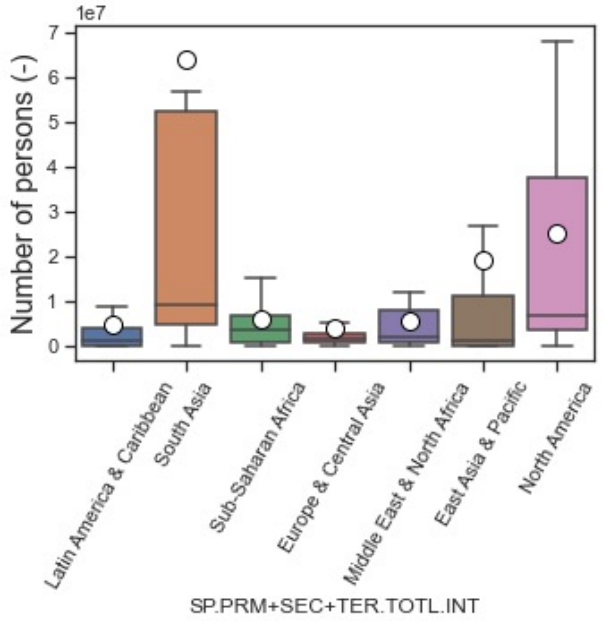
Somme des indicateurs démographiques

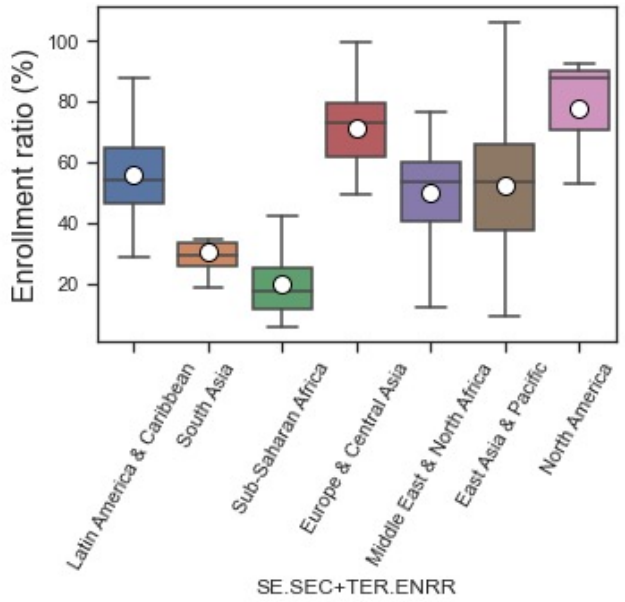
Indicator Code	Indicator Name
NY.GDP.PCAP.CD	GDP per capita (current US$)
SP.PRM+SEC+TER.TOTL.IN	Population of the official age for primary, secondary and tertiary education, both sexes (number)
IT.NET.USER.P2	Internet users (per 100 people)
SE.SEC+TER.ENRR	Gross enrolment mean ratio, tertiary and secondary, both sexes (%)

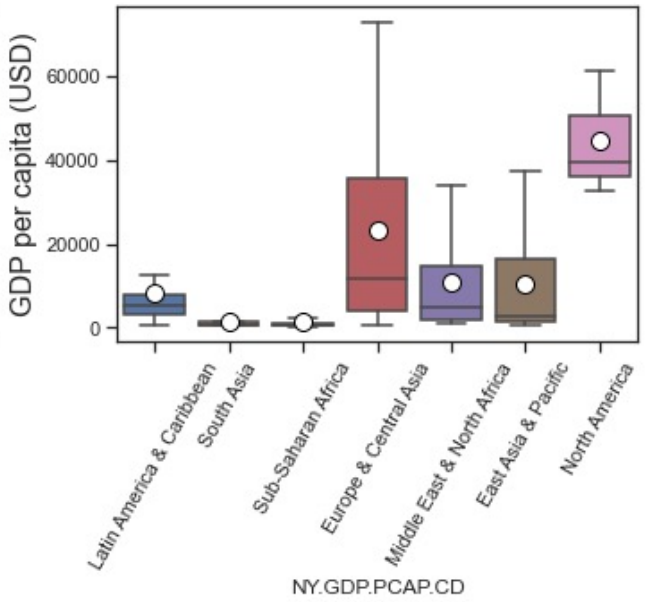
Moyenne des deux indicateurs du taux de scolarisation

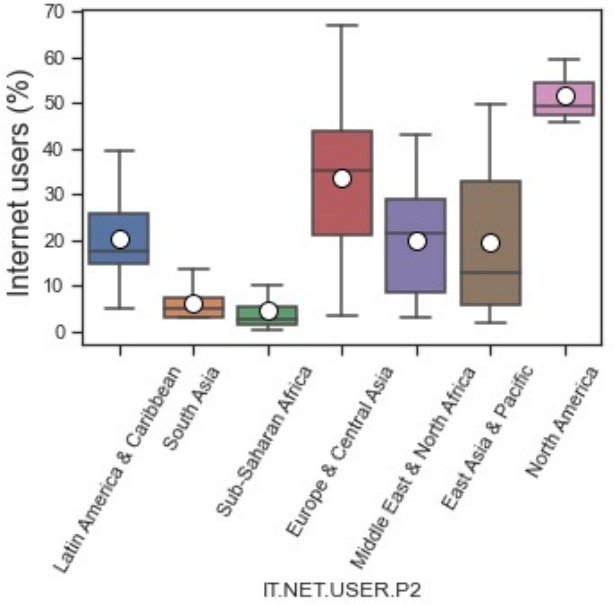
ANALYSE DES NOUVEAUX INDICATEURS

Indicator Code	Indicator Name
NY.GDP.PCAP.CD	GDP per capita (current US$)
SP.PRM+SEC+TER.TOTL.IN	Population of the official age for primary, secondary and tertiary education, both sexes (number)
IT.NET.USER.P2	Internet users (per 100 people)
SE.SEC+TER.ENRR	Gross enrolment mean ratio, tertiary and secondary, both sexes (%)



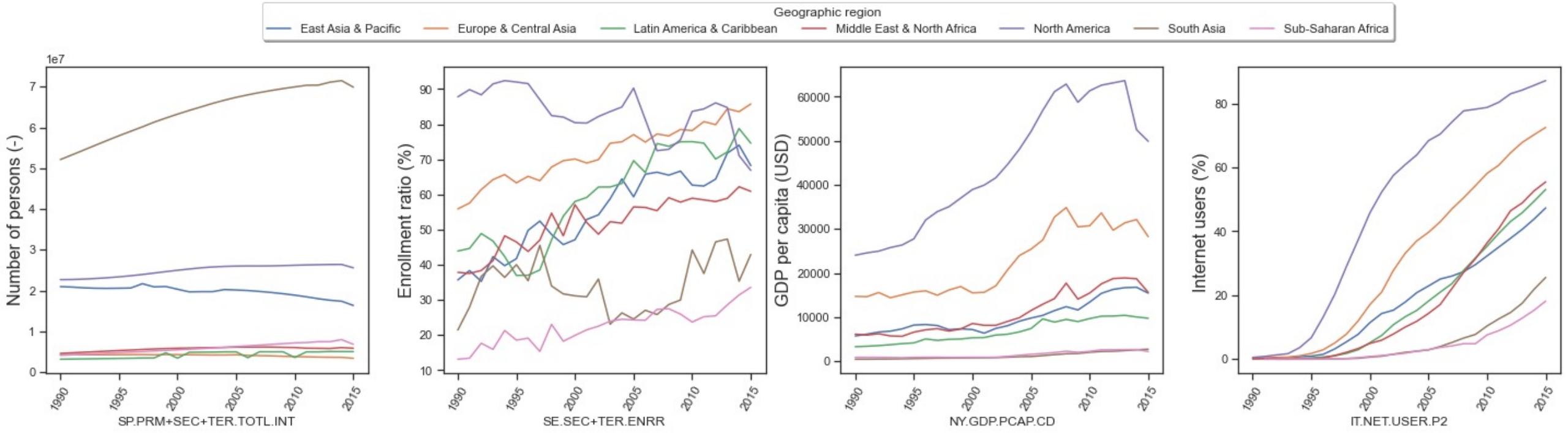






ANALYSE DES NOUVEAUX INDICATEURS

Indicator Code	Indicator Name
NY.GDP.PCAP.CD	GDP per capita (current US$)
SP.PRM+SEC+TER.TOTL.IN	Population of the official age for primary, secondary and tertiary education, both sexes (number)
IT.NET.USER.P2	Internet users (per 100 people)
SE.SEC+TER.ENRR	Gross enrolment mean ratio, tertiary and secondary, both sexes (%)



CRÉATION D'UN SCORE DE SÉLECTION DES PAYS

- Standardisation de chaque indicateur par année :

$$x_{std} = \frac{(x - \mu)}{\sigma}$$

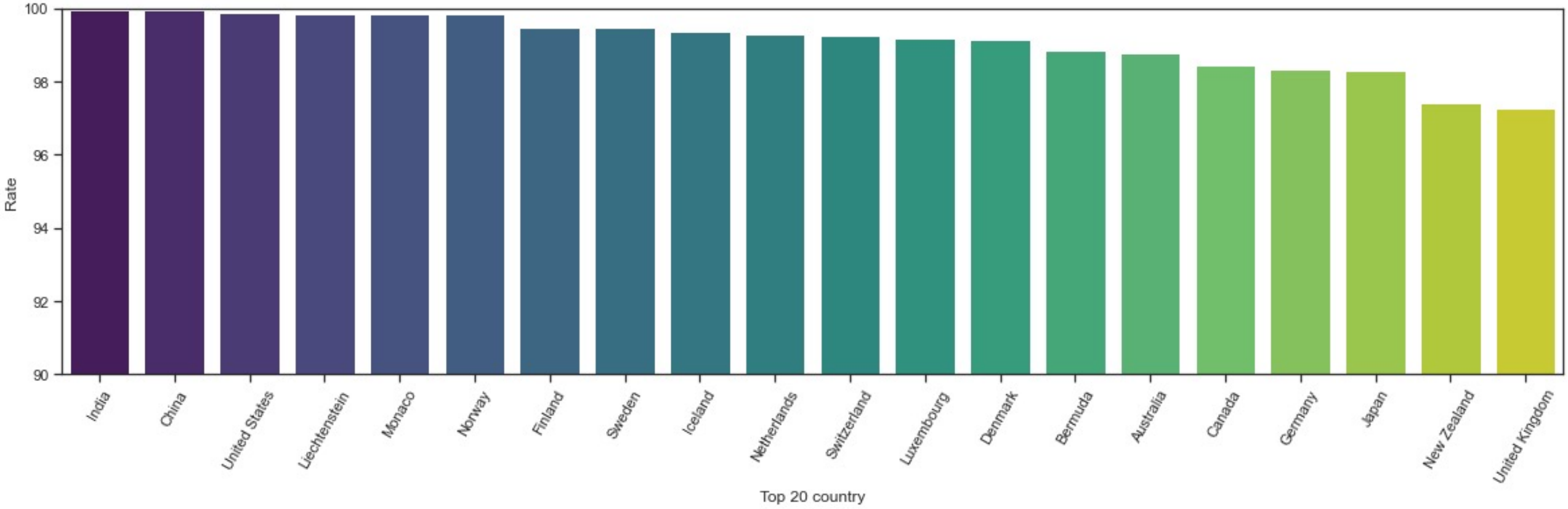
Avec : $\mu = moyenne\ et\ \sigma = l'écart\ type$

- Utilisation de l'équation de la régression logistique :

$$\ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Calcul de p la probabilité de succès d'implantation dans un pays (1 = succès, 0 = échec)

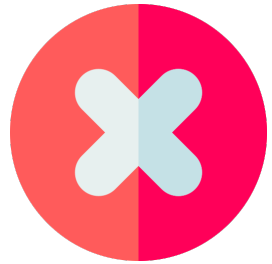
CRÉATION D'UN SCORE DE SÉLECTION DES PAYS



CONCLUSION

Les données sur l'éducation de la banque mondiale permettent elles d'informer le projet d'expansion à l'international ?

- Grande quantité de donnée disponible
- Indicateurs pertinents
- Information sur des pays très différents



- Beaucoup de valeurs manquantes
- Pas d'information directement lié à la formation en ligne
- Nécessite d'interpréter les données

MERCI DE VOTRE ATTENTION.

DESCRIPTION DU JEU DE DONNÉES

EdStatsSeries.csv

- 3665 lignes / 21 colonnes

	Series Code	Topic	Indicator Name	Short definition	Long definition	...	Related source links	Other web links	Related indicators	License Type	Unnamed: 20
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	BAR.NOED.15UP.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 15+ with n...	Percentage of female population age 15+ with n...	...	NaN	NaN	NaN	NaN	NaN
3	BAR.NOED.15UP.ZS	Attainment	Barro-Lee: Percentage of population age 15+ wi...	Percentage of population age 15+ with no educa...	Percentage of population age 15+ with no educa...	...	NaN	NaN	NaN	NaN	NaN
4	BAR.NOED.2024.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 20-24 with...	Percentage of female population age 20-24 with...	...	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
3660	UIS.XUNIT.USCONST.3.FSGOV	Expenditures	Government expenditure per upper secondary stu...	NaN	Average total (current, capital and transfers)...	...	NaN	NaN	NaN	NaN	NaN
3661	UIS.XUNIT.USCONST.4.FSGOV	Expenditures	Government expenditure per post-secondary non-...	NaN	Average total (current, capital and transfers)...	...	NaN	NaN	NaN	NaN	NaN
3662	UIS.XUNIT.USCONST.56.FSGOV	Expenditures	Government expenditure per tertiary student (c...	NaN	Average total (current, capital and transfers)...	...	NaN	NaN	NaN	NaN	NaN
3663	XGDP.23.FSGOV.FDINSTADM.FFD	Expenditures	Government expenditure in secondary institutio...	Total general (local, regional and central) go...	Total general (local, regional and central) go...	...	NaN	NaN	NaN	NaN	NaN
3664	XGDP.56.FSGOV.FDINSTADM.FFD	Expenditures	Government expenditure in tertiary institution...	Total general (local, regional and central) go...	Total general (local, regional and central) go...	...	NaN	NaN	NaN	NaN	NaN

3665 rows × 21 columns

DESCRIPTION DU JEU DE DONNÉES

EdStatsCountry.csv

- 241 lignes / 32 colonnes

	Country Code	Short Name	Table Name	Long Name	2-alpha code	...	Latest agricultural census	Latest industrial data	Latest trade data	Latest water withdrawal data	Unnamed: 31
0	ABW	Aruba	Aruba	Aruba	AW	...	NaN	NaN	2012.0	NaN	NaN
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	...	2013/14	NaN	2012.0	2000	NaN
2	AGO	Angola	Angola	People's Republic of Angola	AO	...	2015	NaN	NaN	2005	NaN
3	ALB	Albania	Albania	Republic of Albania	AL	...	2012	2010.0	2012.0	2006	NaN
4	AND	Andorra	Andorra	Principality of Andorra	AD	...	NaN	NaN	2006.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
236	XKX	Kosovo	Kosovo	Republic of Kosovo	NaN	...	NaN	NaN	NaN	NaN	NaN
237	YEM	Yemen	Yemen, Rep.	Republic of Yemen	YE	...	NaN	2006.0	2012.0	2005	NaN
238	ZAF	South Africa	South Africa	Republic of South Africa	ZA	...	2007	2010.0	2012.0	2000	NaN
239	ZMB	Zambia	Zambia	Republic of Zambia	ZM	...	2010. Population and Housing Census.	NaN	2011.0	2002	NaN
240	ZWE	Zimbabwe	Zimbabwe	Republic of Zimbabwe	ZW	...	NaN	NaN	2012.0	2002	NaN

241 rows × 32 columns

DESCRIPTION DU JEU DE DONNÉES

EdStatsCountry-Series.csv

- 613 lignes / 4 colonnes

	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
...	...	...	...	...
608	ZAF	SP.POP.GROW	Data sources : Statistics South Africa, United...	NaN
609	ZMB	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
610	ZMB	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
611	ZWE	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
612	ZWE	SP.POP.GROW	Data sources: United Nations World Population ...	NaN

DESCRIPTION DU JEU DE DONNÉES

EdStatsFootNotes.csv

- 643638 lignes / 5 colonnes

	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
...	...	...	...	...	...
643633	ZWE	SH.DYN.MORT	YR2007	Uncertainty bound is 91.6 - 109.3	NaN
643634	ZWE	SH.DYN.MORT	YR2014	Uncertainty bound is 54.3 - 76	NaN
643635	ZWE	SH.DYN.MORT	YR2015	Uncertainty bound is 48.3 - 73.3	NaN
643636	ZWE	SH.DYN.MORT	YR2017	5-year average value between 0s and 5s	NaN
643637	ZWE	SP.POP.GROW	YR2017	5-year average value between 0s and 5s	NaN

643638 rows × 5 columns