Analysing data from education systems

DECEMBER 2021

NAJWA MOULINE

Presentation Outline

1. Objectives
2. Dataset Overview
3. Pre-exploratory analysis
4. Conclusions

Objectives

Context

- Service offered - Elearnings: High school and university level online training content
- Interested in international expansion
- World Bank Education Dataset available

Business Problem

- Identify countries with high potential for this expansion.

Mission

- Consult the dataset provided to conclude if it answers the following 3 questions:
- Which countries have a high customer potential?
- For each of these countries, how will this customer potential evolve?
- In which countries should the company operate as a priority?

Methodology

1. Validate the quality of the dataset
2. Describe the information in the dataset
3. Select information relevant to the issue
4. Determining orders of magnitude of statistical indicators

Dataset Overview (1/2)

5 files with common keys, including a main one – 'Data'.

'Data' has 85% missing values.

No duplicates.

Legend

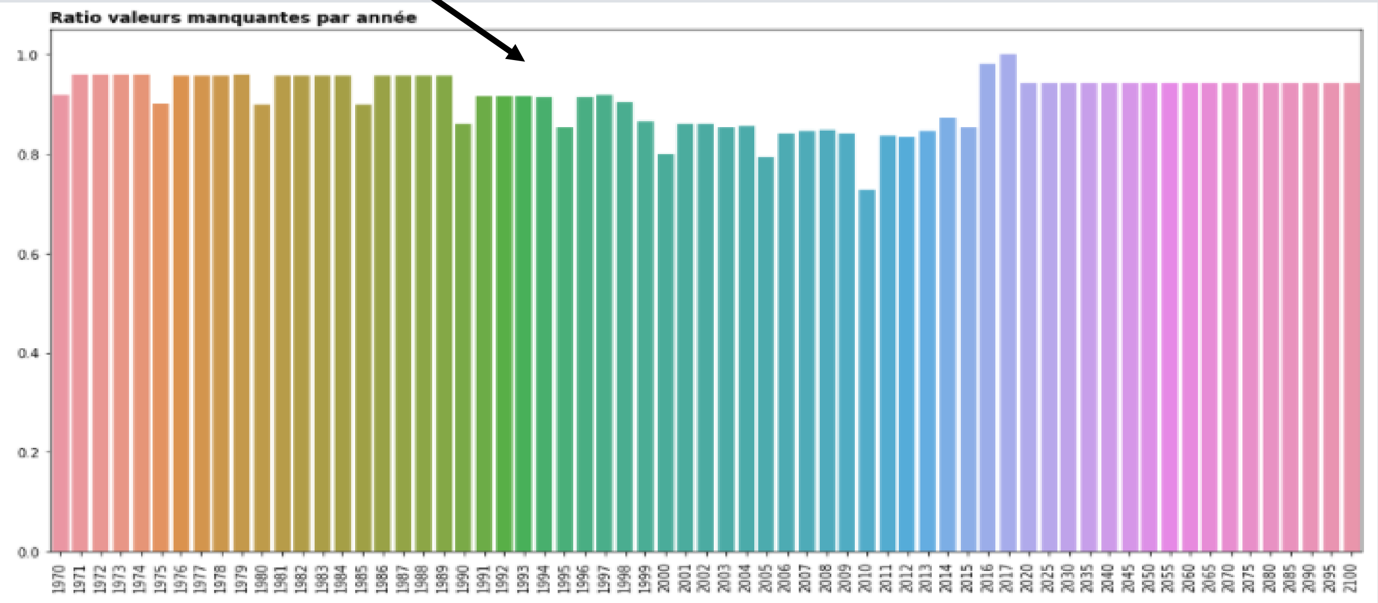
Variable	Nb données	% NaN	Type info	Type stat
Name	241	0%	object	Qual. nom.

Description of the Data file – main file

Data (886930 obs. x 70 columns)

Information by past and future years for each couple (country / indicator)

Country Name	886930	0%	object	Qual. nom.
Country Code*	886930	0%	object	Qual. nom.
Indicator Name	886930	0%	object	Qual. nom.
Indicator Code*	886930	0%	object	Qual. nom.
1970	72288	92%	float64	Quant. cont.
[Others years]	Var.	72-100%	float64	Quant. cont.
Unnamed: 69	0	100%	float64	?



Dataset Overview (2/2)

5 files with common keys, including a main one – 'Data'.

'Data' has 85% missing values.

No duplicates.

Legend

Variable	Nb données	% NaN	Type info	Type stat
Name	241	0%	object	Qual. nom.

Country (241 observations x 32 columns)

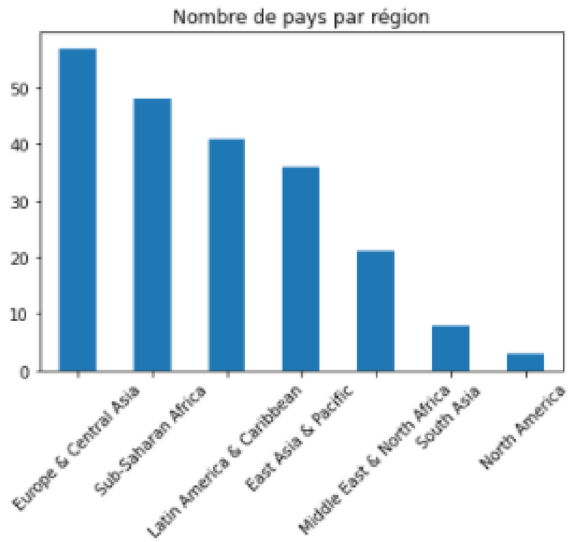
General information on each of the 214 countries

Country Code *	241	0%	object	Qual. nom.
Short Name	241	0%	object	Qual. nom.
Region	214	11%	object	Qual. nom.
Income Group	214	11%	object	Qual. ord.
[...]	Var.	Var.	Var.	Var.
Unnamed: 31	0	100%	float64	?

Series (3665 obs. x 21 columns)

Information on each of the 3665 indicators

Series Code*	3665	0%	object	Qual. nom.
Topic	3665	0%	object	Qual. nom.
Indicator Name	3665	0%	object	Qual. nom.
Long definition	3665	0%	object	Qual. nom.
[...]	Var.	Var.	object	
[5 variables]	0	100%	float64	?



1

*

1

*

Data (886930 obs. x 70 columns) – MAIN TABLE

Information by past and future years for each couple (country / indicator)

Country Code*	886930	0%	object	Qual. nom.
Indicator Code*	886930	0%	object	Qual. nom.
1970	72288	92%	float64	Quant. cont.
[Others years]	Var.	72-100%	float64	Quant. cont.
Unnamed: 69	0	100%	float64	?

Country_Series (613 obs. x 4 columns)

Sources of information for each country/indicator

CountryCode*	613	object	Qual. nom.
SeriesCode*	613	object	Qual. nom.
Description	613	object	Qual. nom.
Unnamed: 3	0	float64	?

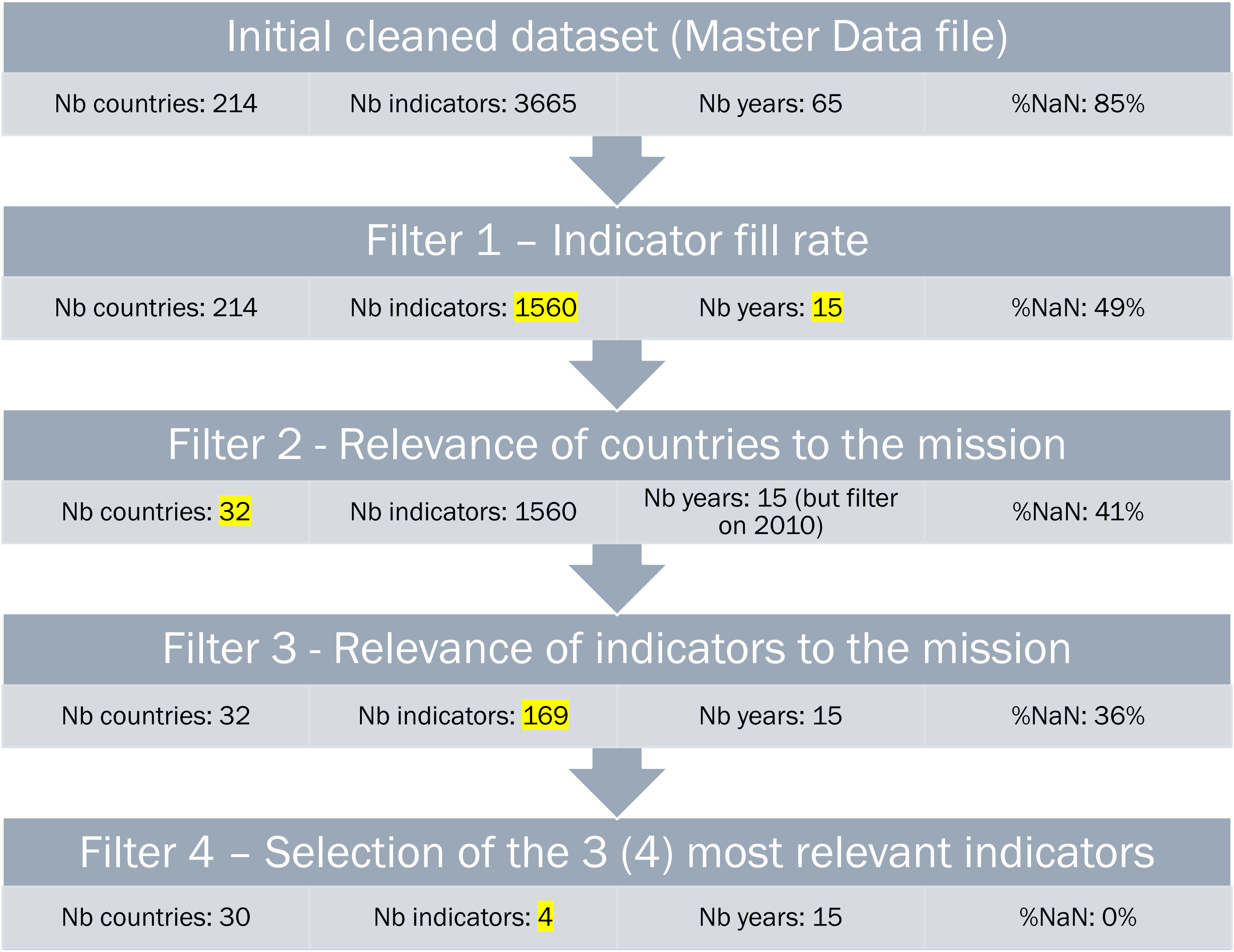
Footnote (643638 obs. x 5 columns)

Sources of information provided for each country/indicator/year

CountryCode*	643638	object	Qual. nom.
SeriesCode*	643638	object	Qual. nom.
Year*	643638	Object	Qual. ord.
Description	643638	object	Qual. nom.
Unnamed: 4	0	float64	Qual. nom.

Pre-exploratory analysis

4-step process



Initial cleaned dataset (Master Data file)

Nb countries: 214	Nb indicators: 3665	Nb years: 65	%NaN: 85%

Filter 1 – Indicator fill rate

Nb countries: 214	Nb indicators: 1560	Nb years: 15	%NaN: 49%

Filter 2 - Relevance of countries to the mission

Nb countries: 32	Nb indicators: 1560	Nb years: 15 (but filter on 2010)	%NaN: 41%

Filter 3 - Relevance of indicators to the mission

Nb countries: 32	Nb indicators: 169	Nb years: 15	%NaN: 36%

Filter 4 – Selection of the 3 (4) most relevant indicators

Nb countries: 30	Nb indicators: 4	Nb years: 15	%NaN: 0%

Pre-exploratory analysis Step 1

Filter of the data from the fill rate (data) of the indicators by indicator/year.

Indicators: 1560

Years: 15

Analysis of the fill rate by indicator and by year

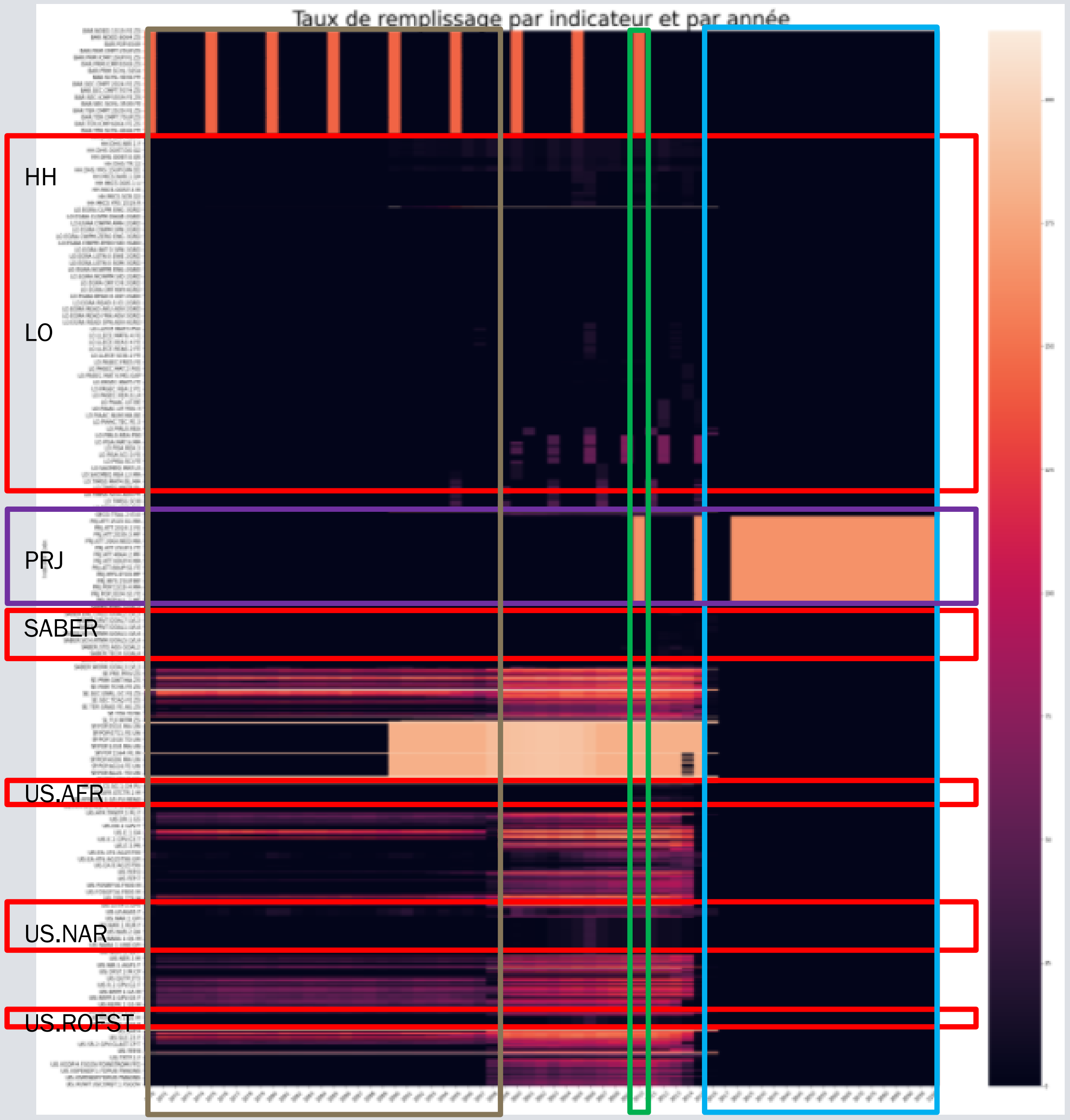
Indicators with almost no data provided
=> Indicators removed

Indicators with only projection data for future years
=> Indicators removed

Years 1970-1998 with little data on indicators
=> Years not considered

Year 2010 with the most data on indicators

Years 2015-2100 with very little or only data on forecast indicators => Years not considered

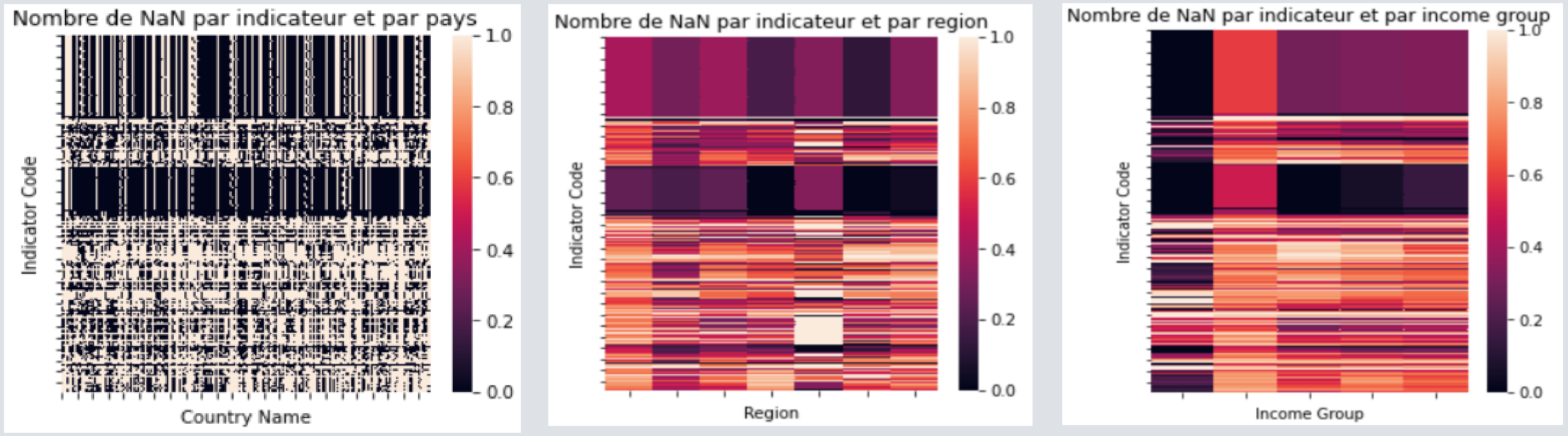


Pre-exploratory analysis Step 2

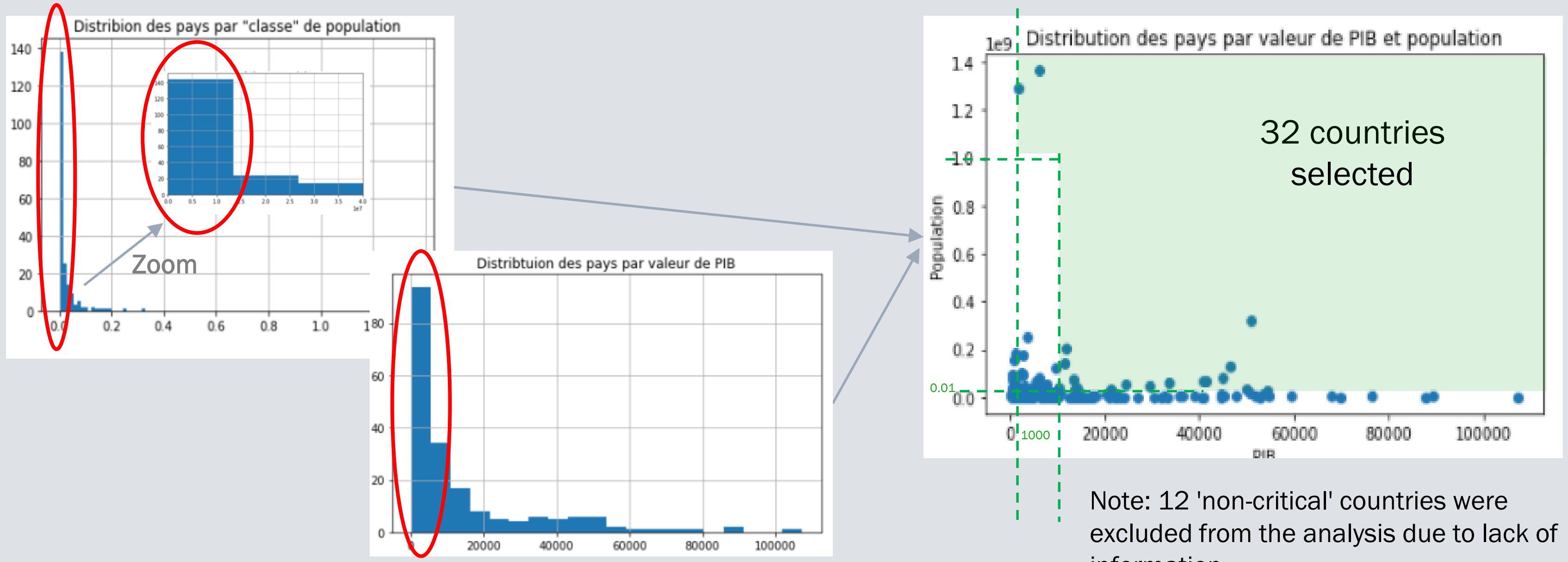
Filter of data from the relevance of countries to the business problem.

Country: 32

Correlation between the number of NaNs and the grouping of countries (in regions, in revenue)? No.



Selection of the most relevant countries – populated and "rich" (year 2014)



32 countries selected

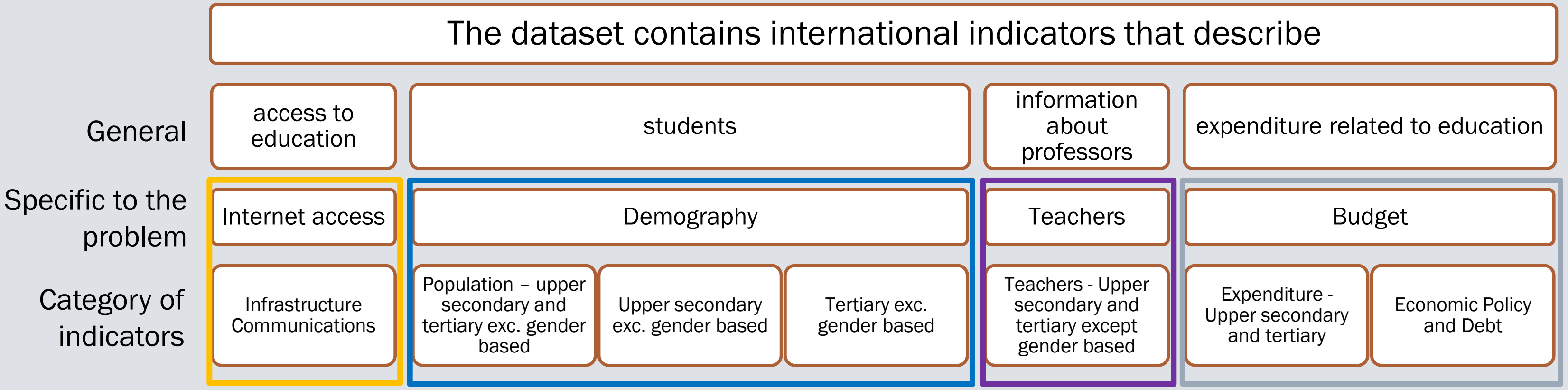
Note: 12 'non-critical' countries were excluded from the analysis due to lack of information.

Pre-exploratory analysis Step 3

Filter of the data from the relevance of the indicators to the business problem.

Indicators: 169

Selection of usable indicators and by relevance of their theme and structure.



The dataset contains international indicators that describe

General

access to education

students

information about professors

expenditure related to education

Specific to the problem

Internet access

Demography

Teachers

Budget

Category of indicators

Infrastructure Communications

Population – upper secondary and tertiary exc. gender based

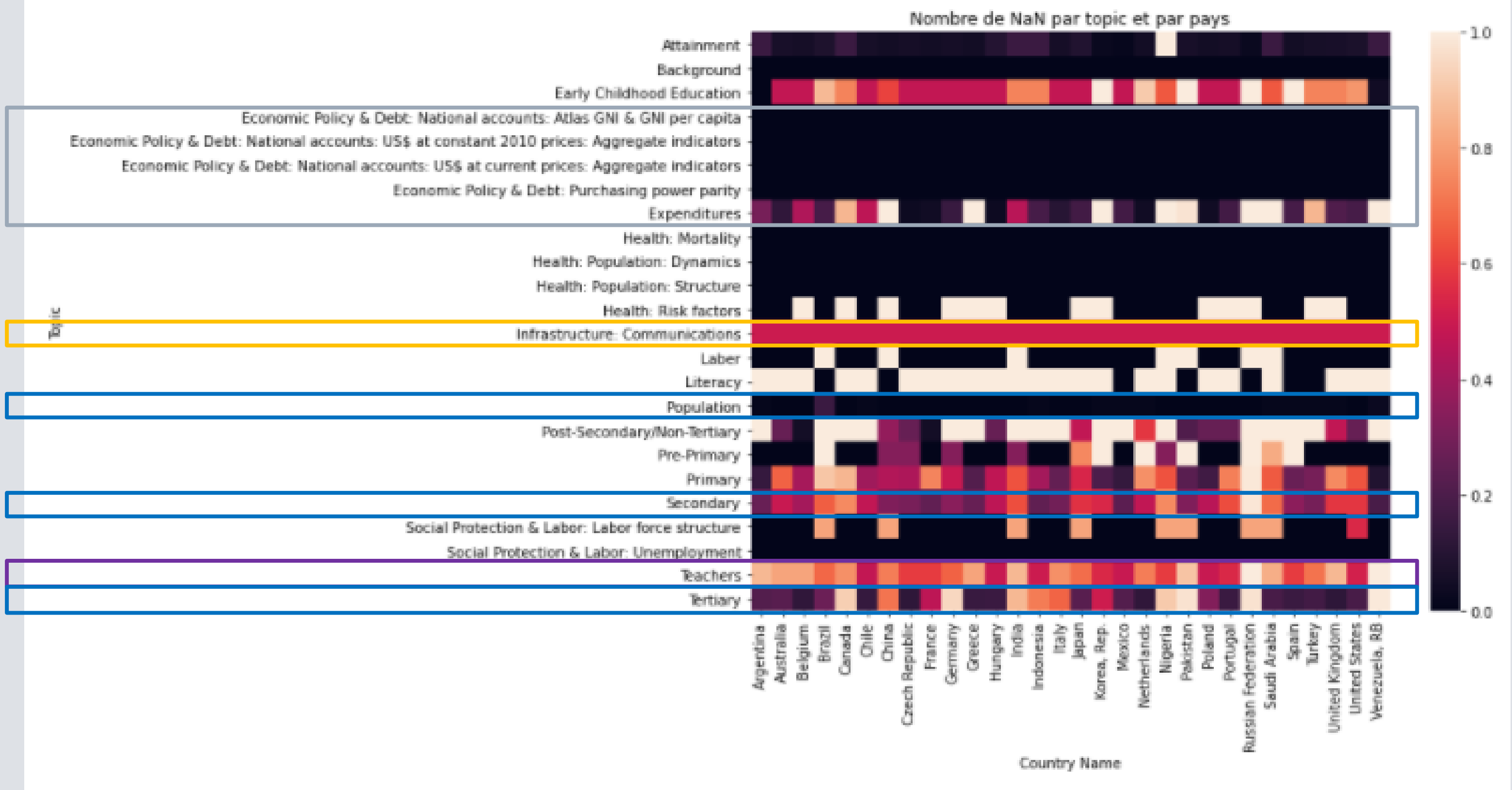
Upper secondary exc. gender based

Tertiary exc. gender based

Teachers - Upper secondary and tertiary except gender based

Expenditure - Upper secondary and tertiary

Economic Policy and Debt

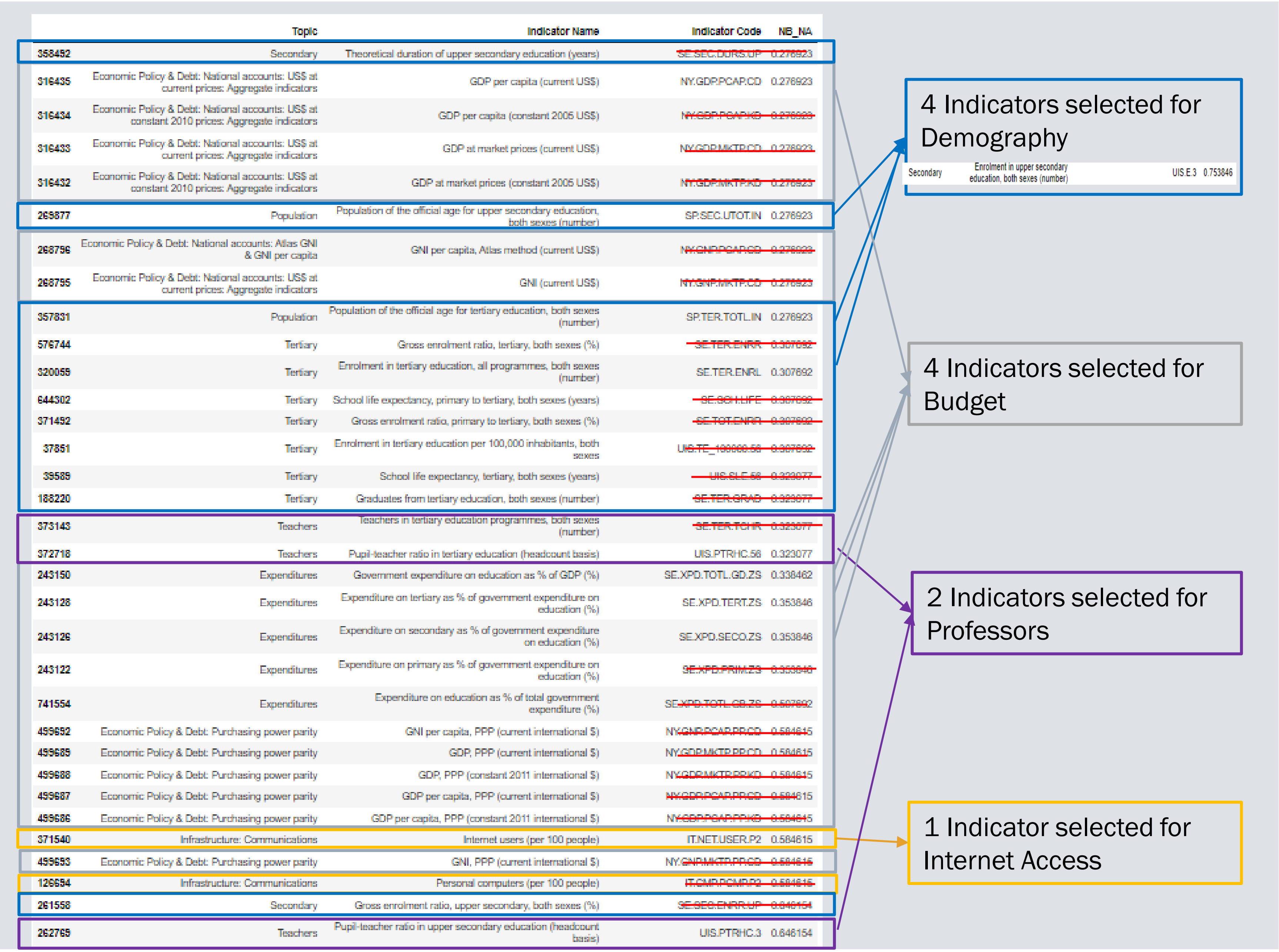


Pre-exploratory analysis Step 4 (1/5)

Data filter to determine the 3 (4) indicators best informed in 2000-2014 and uncorrelated.

Indicators: 11

Selection of indicators by relevance and order of filling.



	Topic	Indicator Name	Indicator Code	NB_NA
358492	Secondary	Theoretical duration of upper secondary education (years)	~~SE.SEC.DURS.UP~~	~~0.276923~~
316435	Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators	GDP per capita (current US$)	NY.GDP.PCAP.CD	0.276923
316434	Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators	GDP per capita (constant 2005 US$)	~~NY.GDP.PCAP.KD~~	~~0.276923~~
316433	Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators	GDP at market prices (current US$)	NY.~~GDP.MKTP.CD~~	~~0.276923~~
316432	Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators	GDP at market prices (constant 2005 US$)	NY.~~GDP.MKTP.KD~~	~~0.276923~~
269877	Population	Population of the official age for upper secondary education, both sexes (number)	SP.SEC.UTOT.IN	0.276923
268796	Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita	GNI per capita, Atlas method (current US$)	~~NY.GNP.PCAP.CD~~	~~0.276923~~
268795	Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators	GNI (current US$)	~~NY.GNP.MKTP.CD~~	~~0.276923~~
357831	Population	Population of the official age for tertiary education, both sexes (number)	SP.TER.TOTL.IN	0.276923
576744	Tertiary	Gross enrolment ratio, tertiary, both sexes (%)	~~SE.TER.ENRR~~	~~0.307692~~
320059	Tertiary	Enrolment in tertiary education, all programmes, both sexes (number)	SE.TER.ENRL	0.307692
644302	Tertiary	School life expectancy, primary to tertiary, both sexes (years)	~~SE.SCH.LIFE~~	~~0.307692~~
371452	Tertiary	Gross enrolment ratio, primary to tertiary, both sexes (%)	~~SE.TOT.ENRR~~	~~0.307692~~
37851	Tertiary	Enrolment in tertiary education per 100,000 inhabitants, both sexes	U~~IS.TE_100000.56~~	~~0.307692~~
39589	Tertiary	School life expectancy, tertiary, both sexes (years)	~~UIS.SLE.56~~	~~0.323077~~
188220	Tertiary	Graduates from tertiary education, both sexes (number)	~~SE.TER.GRAD~~	~~0.323077~~
373143	Teachers	Teachers in tertiary education programmes, both sexes (number)	~~SE.TER.TCHR~~	~~0.323077~~
372718	Teachers	Pupil-teacher ratio in tertiary education (headcount basis)	UIS.PTRHC.56	0.323077
243150	Expenditures	Government expenditure on education as % of GDP (%)	SE.XPD.TOTL.GD.ZS	0.338462
243128	Expenditures	Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	0.353846
243126	Expenditures	Expenditure on secondary as % of government expenditure on education (%)	SE.XPD.SECO.ZS	0.353846
243122	Expenditures	Expenditure on primary as % of government expenditure on education (%)	~~SE.XPD.PRIM.ZS~~	~~0.353846~~
741554	Expenditures	Expenditure on education as % of total government expenditure (%)	SE.~~XPD.TOTL.GB.ZS~~	~~0.507692~~
499692	Economic Policy & Debt: Purchasing power parity	GNI per capita, PPP (current international $)	NY.~~GNP.PCAP.PP.CD~~	~~0.584615~~
499689	Economic Policy & Debt: Purchasing power parity	GDP, PPP (current international $)	NY.~~GDP.MKTP.PP.CD~~	~~0.584615~~
499688	Economic Policy & Debt: Purchasing power parity	GDP, PPP (constant 2011 international $)	NY.~~GDP.MKTP.PP.KD~~	~~0.584615~~
499687	Economic Policy & Debt: Purchasing power parity	GDP per capita, PPP (current international $)	~~NY.GDP.PCAP.PP.CD~~	~~0.584615~~
499686	Economic Policy & Debt: Purchasing power parity	GDP per capita, PPP (constant 2011 international $)	NY.~~GDP.PCAP.PP.KD~~	~~0.584615~~
371540	Infrastructure: Communications	Internet users (per 100 people)	IT.NET.USER.P2	0.584615
499693	Economic Policy & Debt: Purchasing power parity	GNI, PPP (current international $)	NY.~~GNP.MKTP.PP.CD~~	~~0.584615~~
126694	Infrastructure: Communications	Personal computers (per 100 people)	~~IT.CMP.PCMP.P2~~	~~0.584615~~
261558	Secondary	Gross enrolment ratio, upper secondary, both sexes (%)	~~SE.SEC.ENRR.UP~~	~~0.646154~~
262769	Teachers	Pupil-teacher ratio in upper secondary education (headcount basis)	UIS.PTRHC.3	0.646154

4 Indicators selected for Demography

Secondary	Enrolment in upper secondary education, both sexes (number)	UIS.E.3	0.753846

4 Indicators selected for Budget

2 Indicators selected for Professors

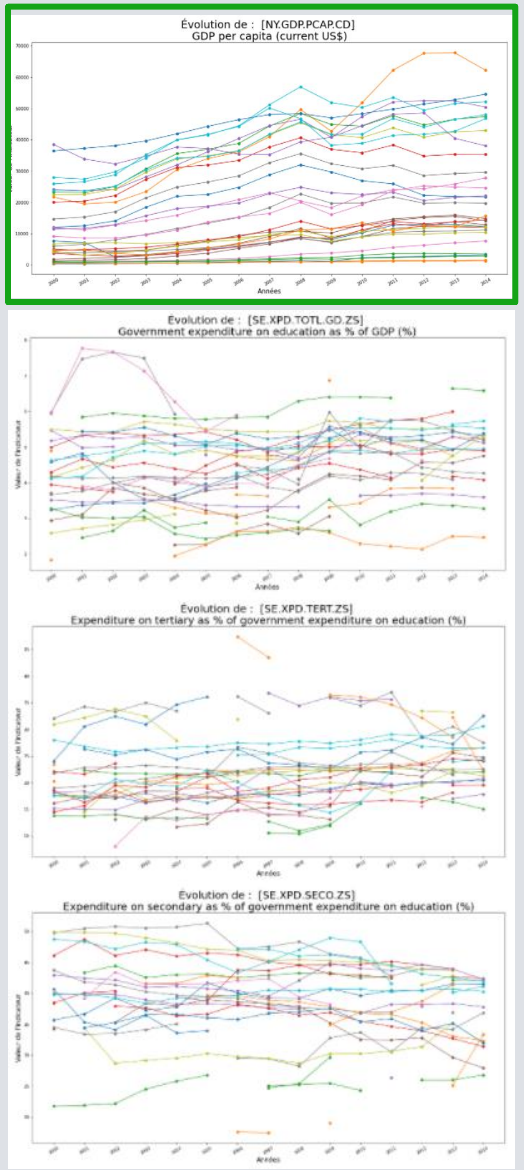
1 Indicator selected for Internet Access

Pre-exploratory analysis Step 4 (2/5)

Filters data to determine the most relevant budget indicator(s).

Analysis (years 2000-2014) and correlation (2014) of budget indicators

Evolution of the 4 indicators



Évolution de : [NY.GDP.PCAP.CD]
GDP per capita (current US$)

Evolution de : [SE.XPD.TOTL.GD.ZS]
Government expenditure on education as % of GDP (%)

Évolution de : [SE.XPD.TERT.ZS]
Expenditure on tertiary as % of government expenditure on education (%)

Evolution de : [SE.XPD.SECO.ZS]
Expenditure on secondary as % of government expenditure on education (%)

Indicator relevance criteria:

- Homogeneous filling over 2000-2014
- Varies from country to country
- **Growing with time**
- Covering both high school and uni
- *Not correlated with other selected indicators*

Selected indicator(s):

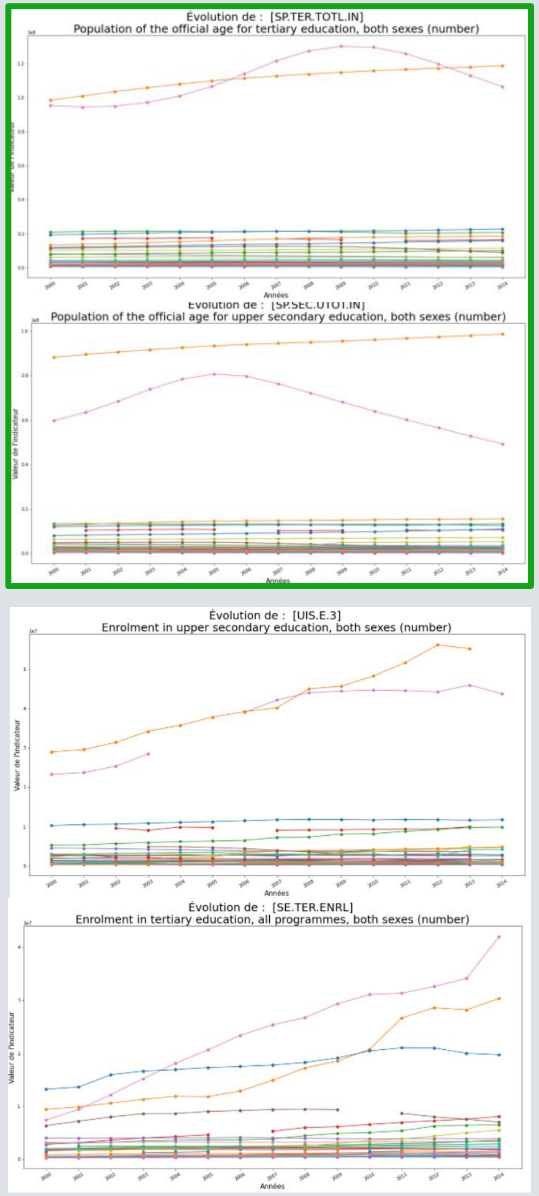
- NY.GDP.PCAP.CD (GDP per capita)

Pre-exploratory analysis Step 4 (3/5)

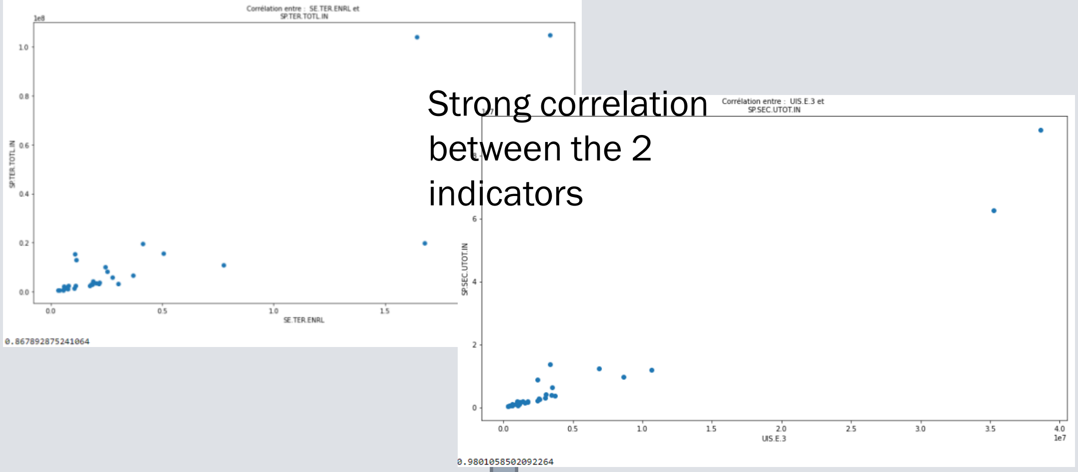
Filters data to determine the most relevant demographic indicator(s).

Analysis (years 2000-2014) and correlation (2014) of demography indicators

Evolution of the 4 indicators



Indicator relevance criteria:
- Homogeneous filling over 2000-2014
- Varies from country to country
- **Growing with time**
- Covering both high school and uni
- *Not correlated with other selected indicators*



Strong correlation between the 2 indicators

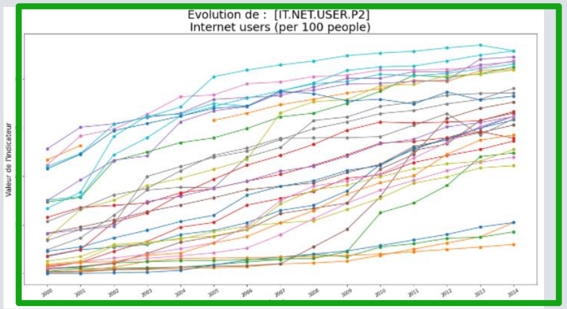
Selected indicator(s):
- SP.SEC.UTOT.IN (Pop of official secondary)
- SP.TER.TOTL.IN (Pop of official tertiary)

Pre-exploratory analysis Step 4 (4/5)

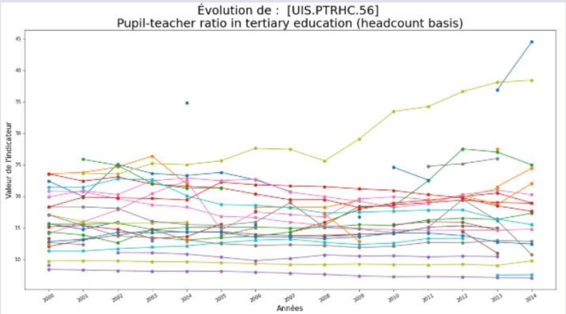
Filters data to determine the most relevant indicator(s) for other categories.

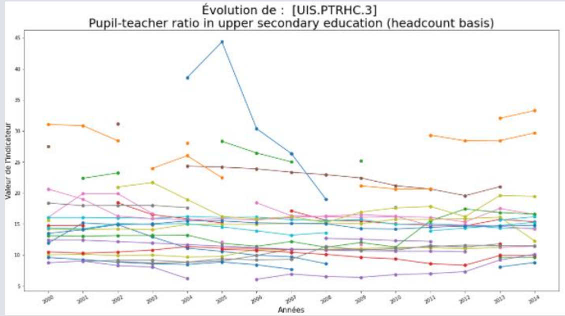
Analysis (years 2000-2014) and correlation (2014) of other indicators

Evolution of the indicator Infrastructure



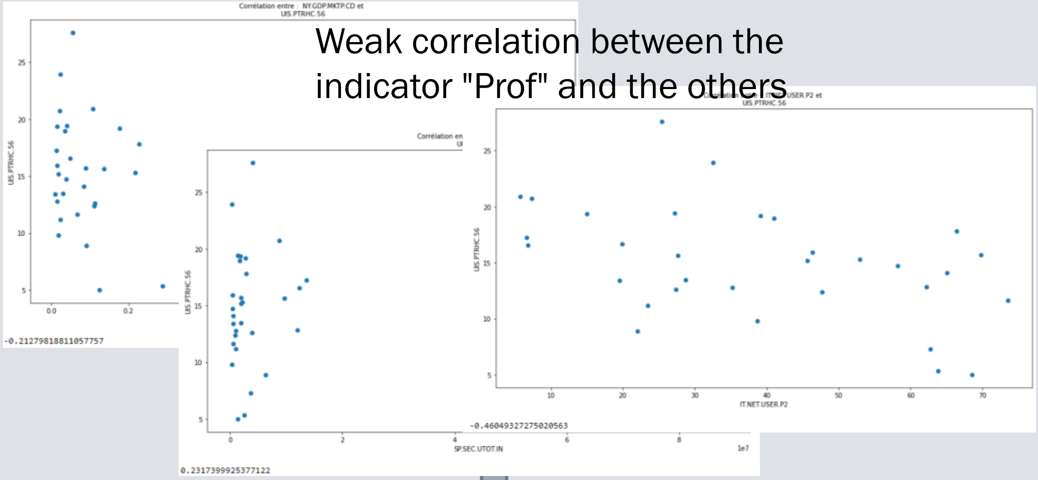
Evolution of the 2 indicators Teachers





Indicator relevance criteria:

- Homogeneous filling over 2000-2014
- Varies from country to country
- **Growing with time**
- Covering both high school and uni
- *Not correlated with other selected indicators*



```
-0.21279818811057757
```

```
0.2317399925377122
```

```
-0.46049327275020563
```

Weak correlation between the indicator "Prof" and the others

Selected indicator(s):

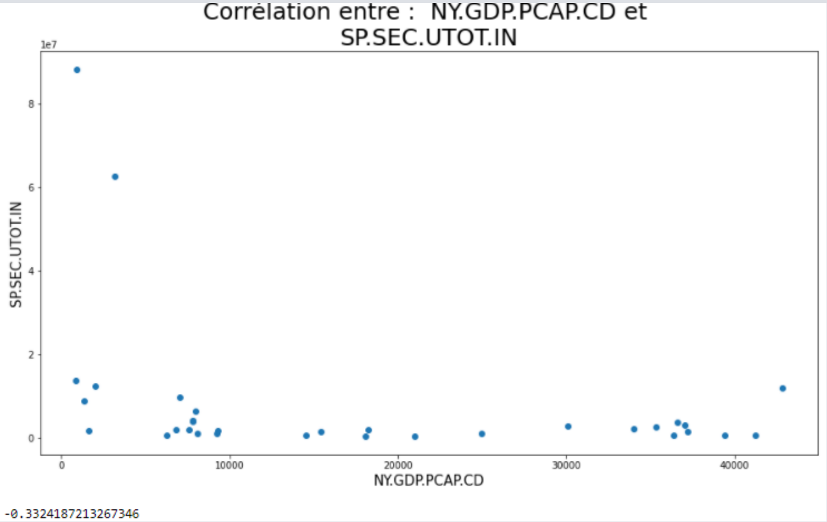
- IT.NET.USER.P2 (Internet per 100 users)
- No indicators kept for "Professors"

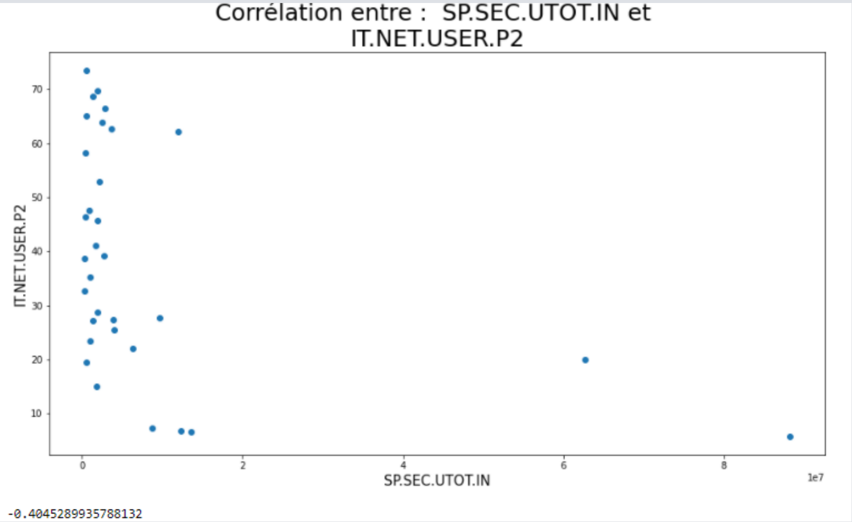
Pre-exploratory analysis Step 4 (5/5)

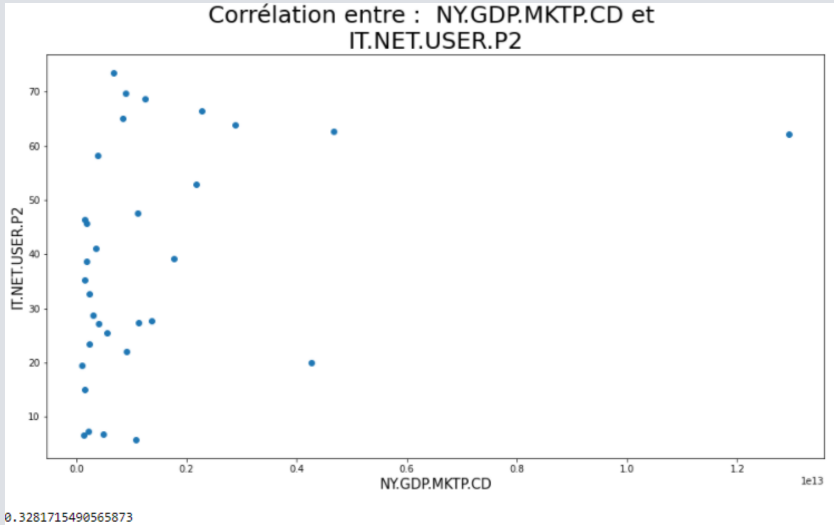
Data filter to determine the 3 (4) indicators best informed in 2000-2014 and uncorrelated.

Indicators: 4

Correlations (2014) between selected indicators







Slight correlation between indicators:

- Budget and Demography
- Infrastructure and Demography
- Infrastructure and Budget

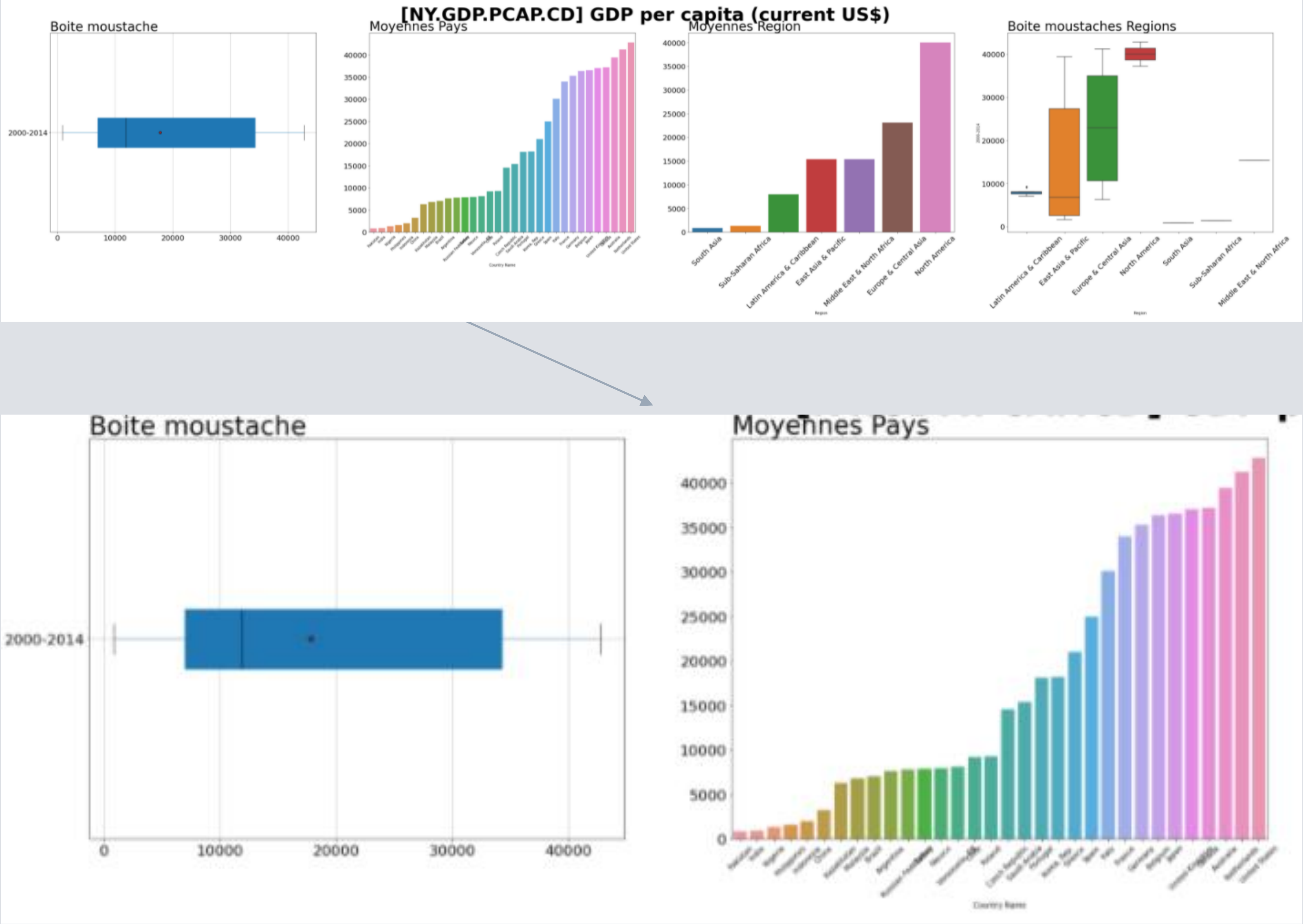
Selected indicator(s): The 4 indicators previously chosen

Conclusions Order of magnitude (1/2)

Order of magnitude of selected indicators – overall, by country and region.

Great disparity of orders of magnitude.

Orders of magnitude of the statistical indicators for the different geographical areas and countries of the world (mean/median/standard deviation by country/geographical block) (average years 2000-2014)

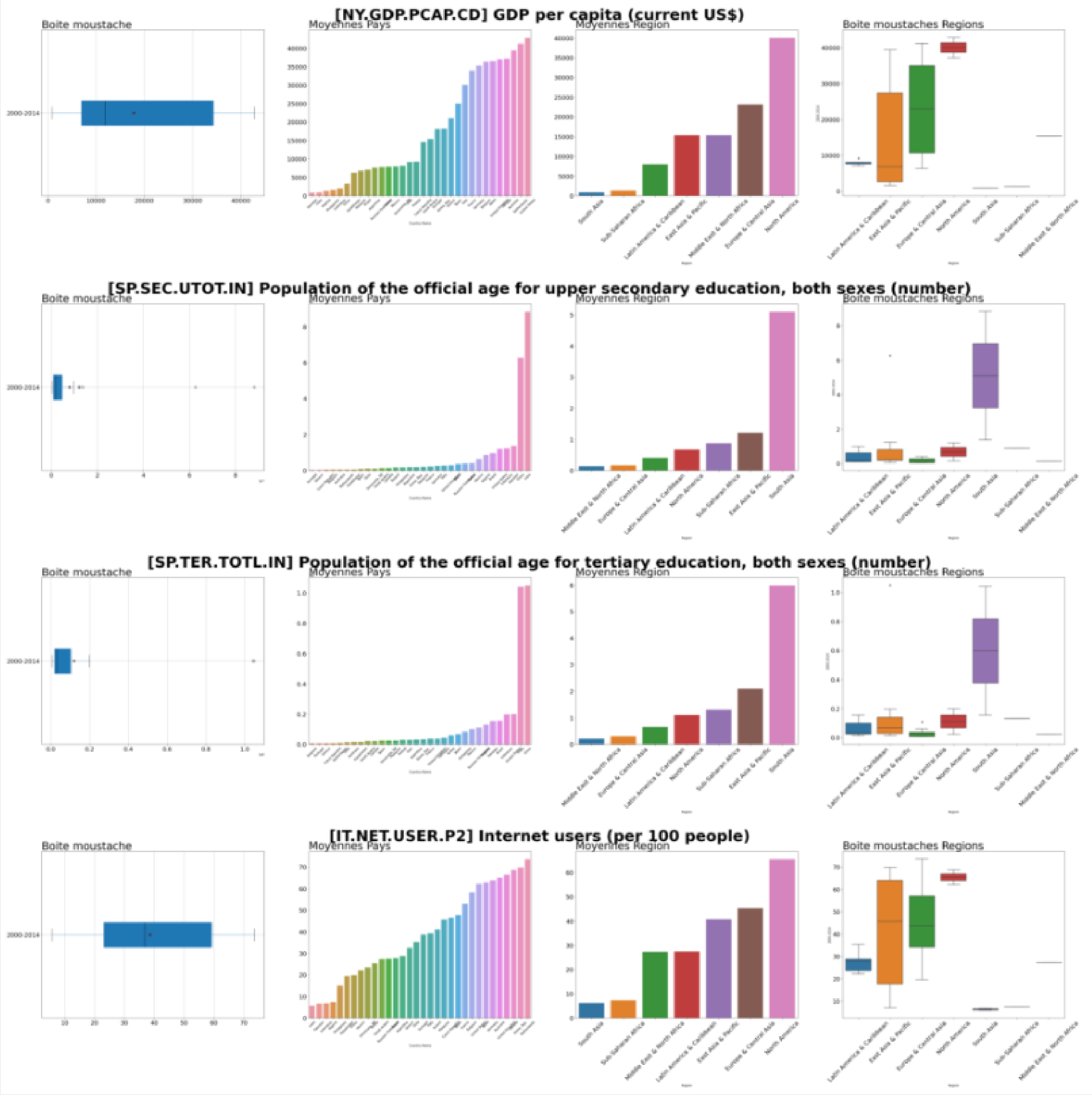


Conclusions Order of magnitude (2/2)

Order of magnitude of selected indicators – overall, by country and region.

Great disparity of orders of magnitude.

Orders of magnitude of the statistical indicators for the different geographical areas and countries of the world (mean/median/standard deviation by country/geographical block) (average years 2000-2014)



[NY.GDP.PCAP.CD] GDP per capita (current US$)

[SP.SEC.UTOT.IN] Population of the official age for upper secondary 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)

"Budget" dominated by the countries of North America and Europe.

Number of high school and university students dominated by the countries of South Asia.

Non-aberrant outliers. China and India have a larger population.

"Infrastructure" dominated by the countries of North America and Europe.

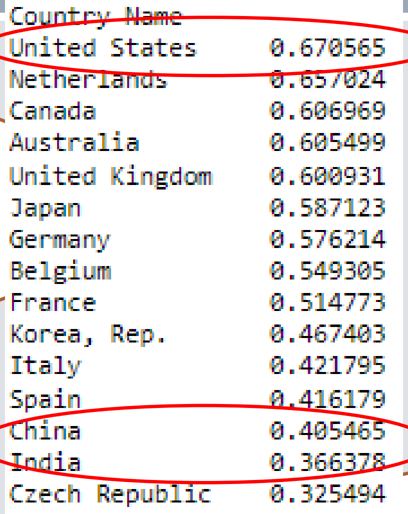
Conclusions Potential countries

Identification of countries relevant to the business problem

Countries with high potential, customer evolution and countries to be considered as a priority (Mid-years 2000-2014)

Which countries have a high customer potential?

Top 15

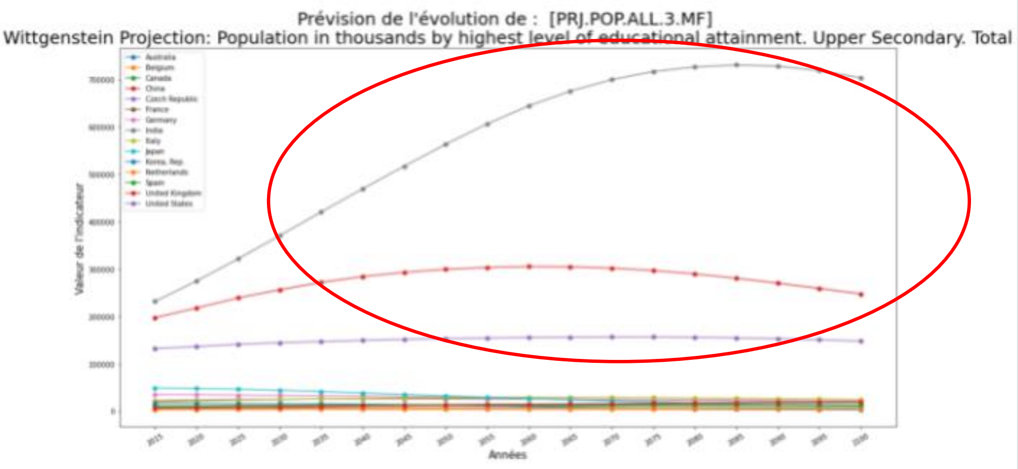


```
Country Name
United States       0.670565
Netherlands         0.657024
Canada              0.606969
Australia           0.605499
United Kingdom      0.600931
Japan               0.587123
Germany             0.576214
Belgium             0.549305
France              0.514773
Korea, Rep.         0.467403
Italy               0.421795
Spain               0.416179
China               0.405465
India               0.366378
Czech Republic      0.325494
```

In which countries to operate as a priority?

US, China and India

For these countries, what will be the evolution of customers?



Conclusions Relevance of the dataset

This dataset is a good starting point for identifying potential countries, but insufficient to confirm them and confirm their priority.

Relevance and limitation of the dataset

Relevance of the dataset

- Countries and regions all represented
- Lots of data related to education
- Specified and known sources

Dataset limitations

- Most recent values are from 2014
- Some interesting but unusable indicators
- No specific indicators:
  - the language of instruction
  - the country's policy: political stability, taxes, ...
  - the local competition, ...

Thank you for your attention!