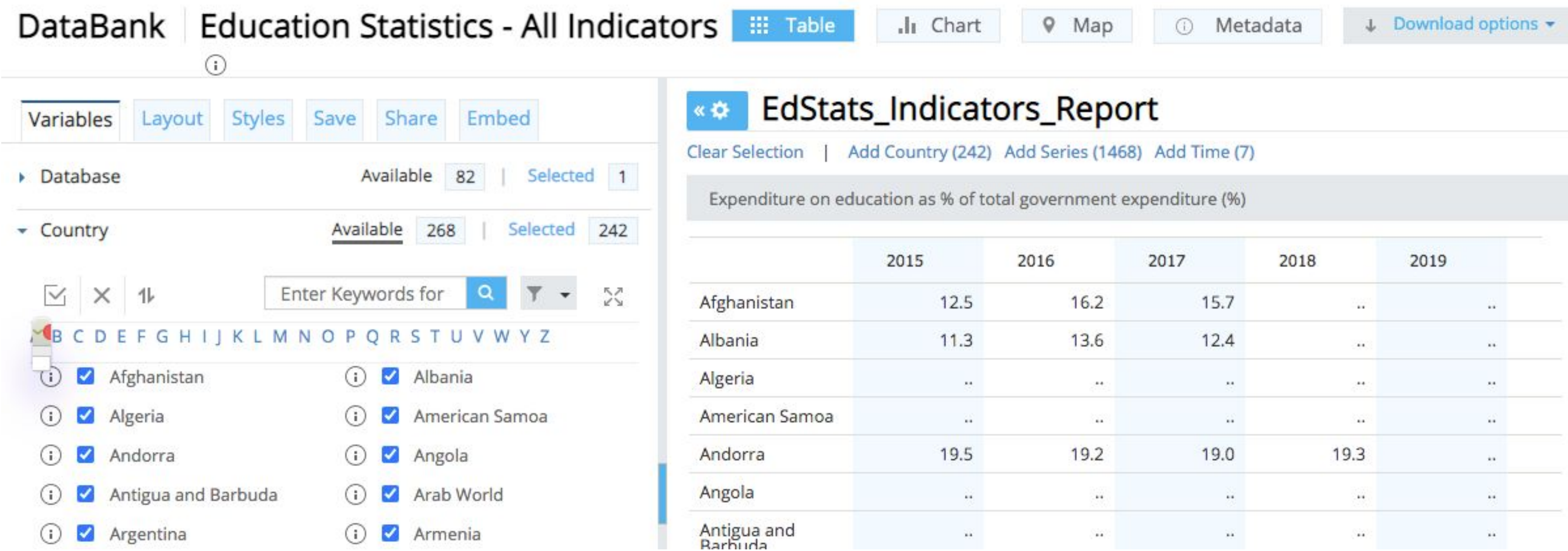
Sommaire

1. Présentation du jeu de données
2. Présentation de l’analyse pré-exploratoire
3. Conclusion sur la pertinence d’utiliser ce jeu de données
4. Questions

Présentation du jeu de données

- Données de la banque mondiale: https://databank.worldbank.org/



DataBank
Education Statistics - All Indicators
Table
Chart
Map
Metadata
Download options
Variables
Layout
Styles
Save
Share
Embed
Enter Keywords for
B C D E F G H I J K L M N O P Q R S T U V W Y Z
EdStats_Indicators_Report
Clear Selection | Add Country (242) Add Series (1468) Add Time (7)

Database Available 82 | Selected 1

Country Available 268 | Selected 242

☑ Afghanistan ☑ Albania
☑ Algeria ☑ American Samoa
☑ Andorra ☑ Angola
☑ Antigua and Barbuda ☑ Arab World
☑ Argentina ☑ Armenia

Expenditure on education as % of total government expenditure (%)

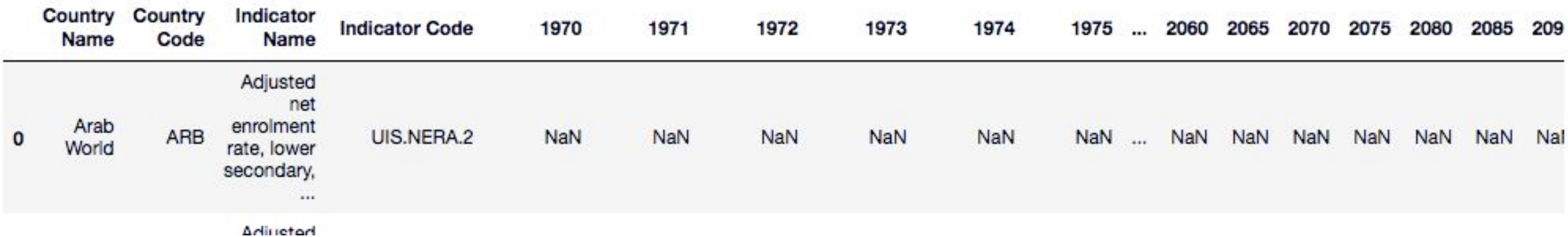
	2015	2016	2017	2018	2019
Afghanistan	12.5	16.2	15.7	..	..
Albania	11.3	13.6	12.4	..	..
Algeria	..	..	..	..	..
American Samoa	..	..	..	..	..
Andorra	19.5	19.2	19.0	19.3	..
Angola	..	..	..	..	..
Antigua and Barbuda	..	..	..	..	..

Présentation du jeu de données

- Données de la banque mondiale: https://databank.worldbank.org/
- 5 tables: EdStatsData, EdStatsSeries, EdStatsCountry, EdStatsCountrySeries et EdStatsFootNote
- Exploration de chacune de ces tables

Présentation du jeu de données

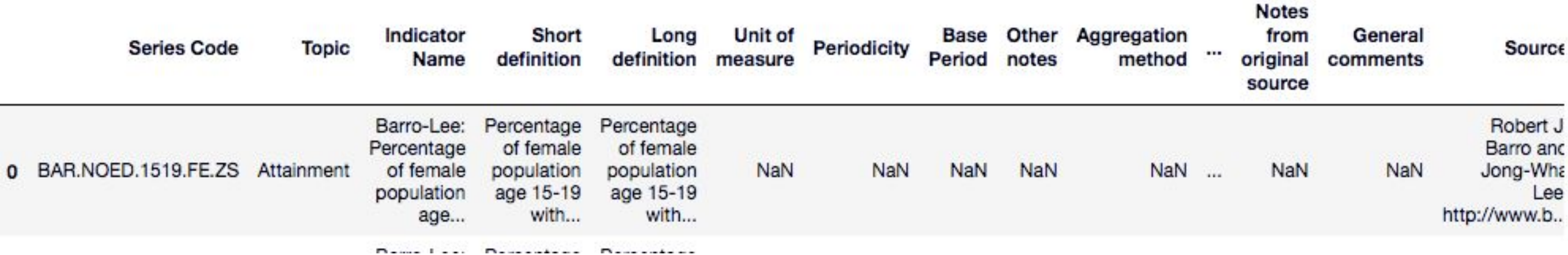
- EdStatsData



	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	209
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Na
			Adjusted															

Présentation du jeu de données

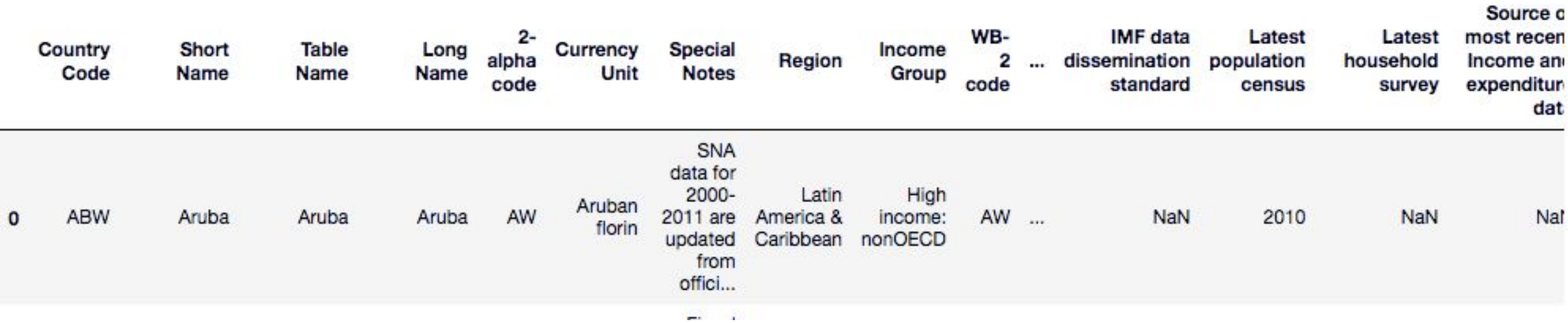
- EdStatsSeries



	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...	Notes from original source	General comments	Source
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	...	NaN	NaN	Robert J Barro and Jong-Wha Lee http://www.b...

Présentation du jeu de données

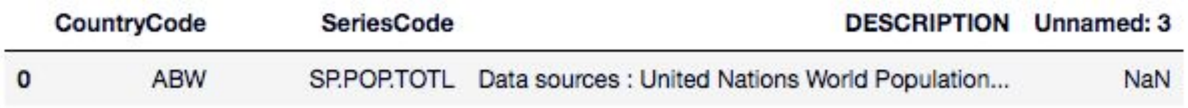
- EdStatsCountry



	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	IMF data dissemination standard	Latest population census	Latest household survey	Source o most recen Income an expenditur dat
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD	AW	...	NaN	2010	NaN	Nal

Présentation du jeu de données

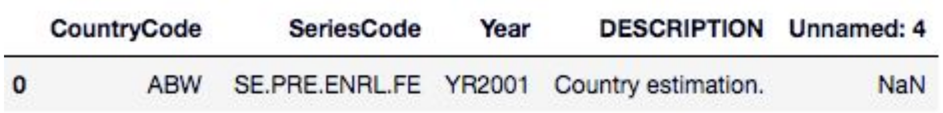
- EdStatscountrySeries



	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN

Présentation du jeu de données

- EdStatsFootNote



	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN

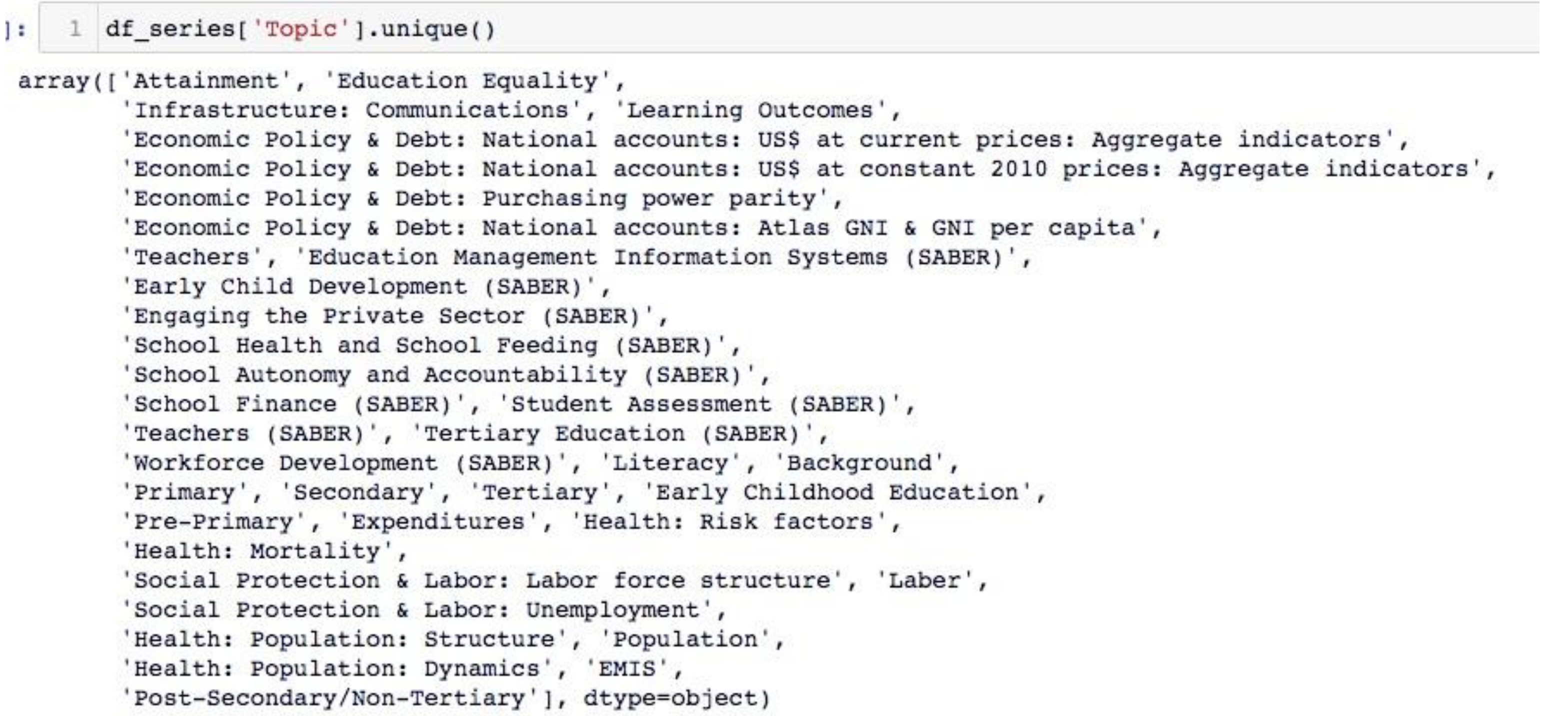
Analyse pré-exploratoire

- Nettoyage du dataset
  - Indicateurs
  - Filtrage pays/régions
  - Années 1: données passées vs. données prospectives
  - Sélection du dataset pays/données passées
  - Années 2: sélection des années 2000-2016
  - Sélection par la population
  - Sélection par le taux de remplissage
  - Sélection par les corrélations
- Statistiques
  - Statistiques pays
  - Statistiques régions
- Scoring et sélection de pays avec du potentiel

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



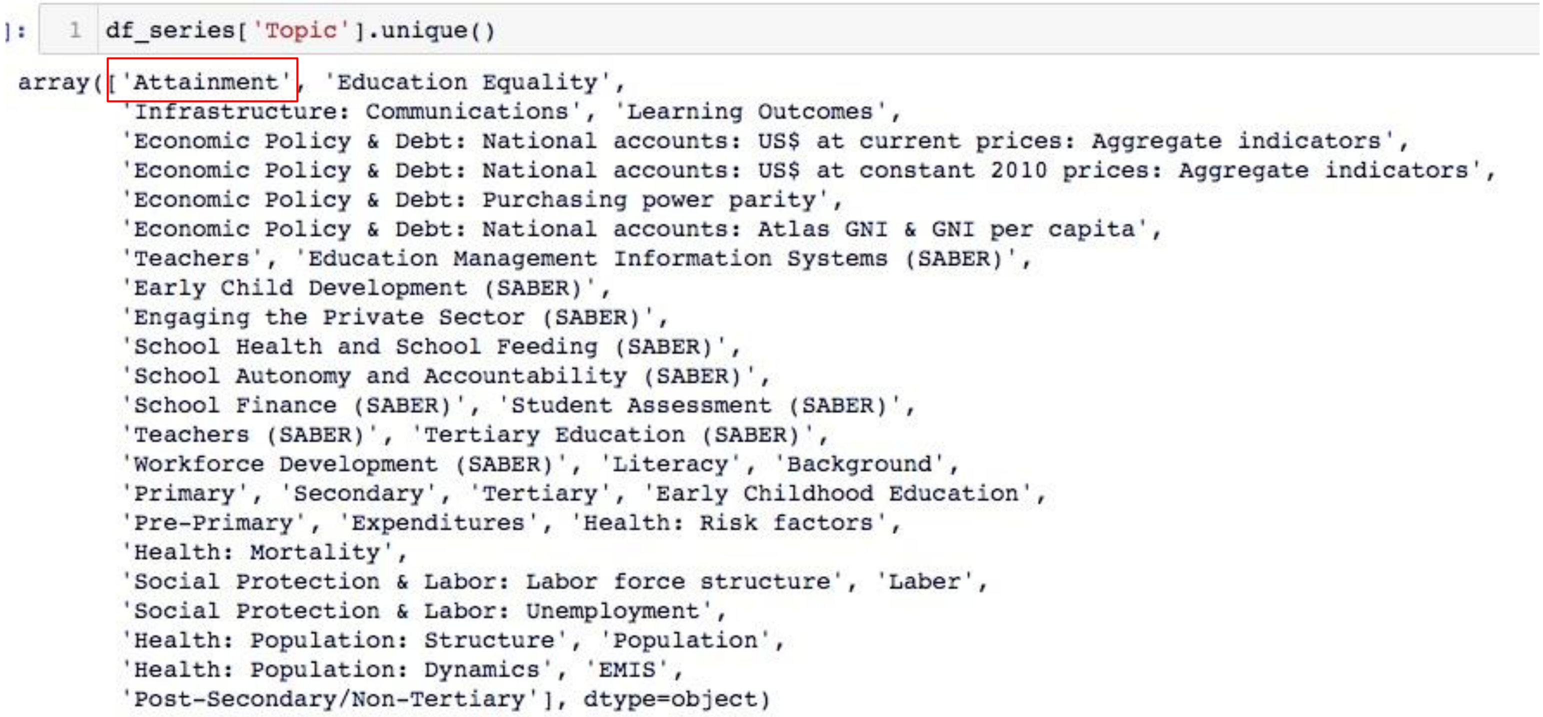
```
df_series['Topic'].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



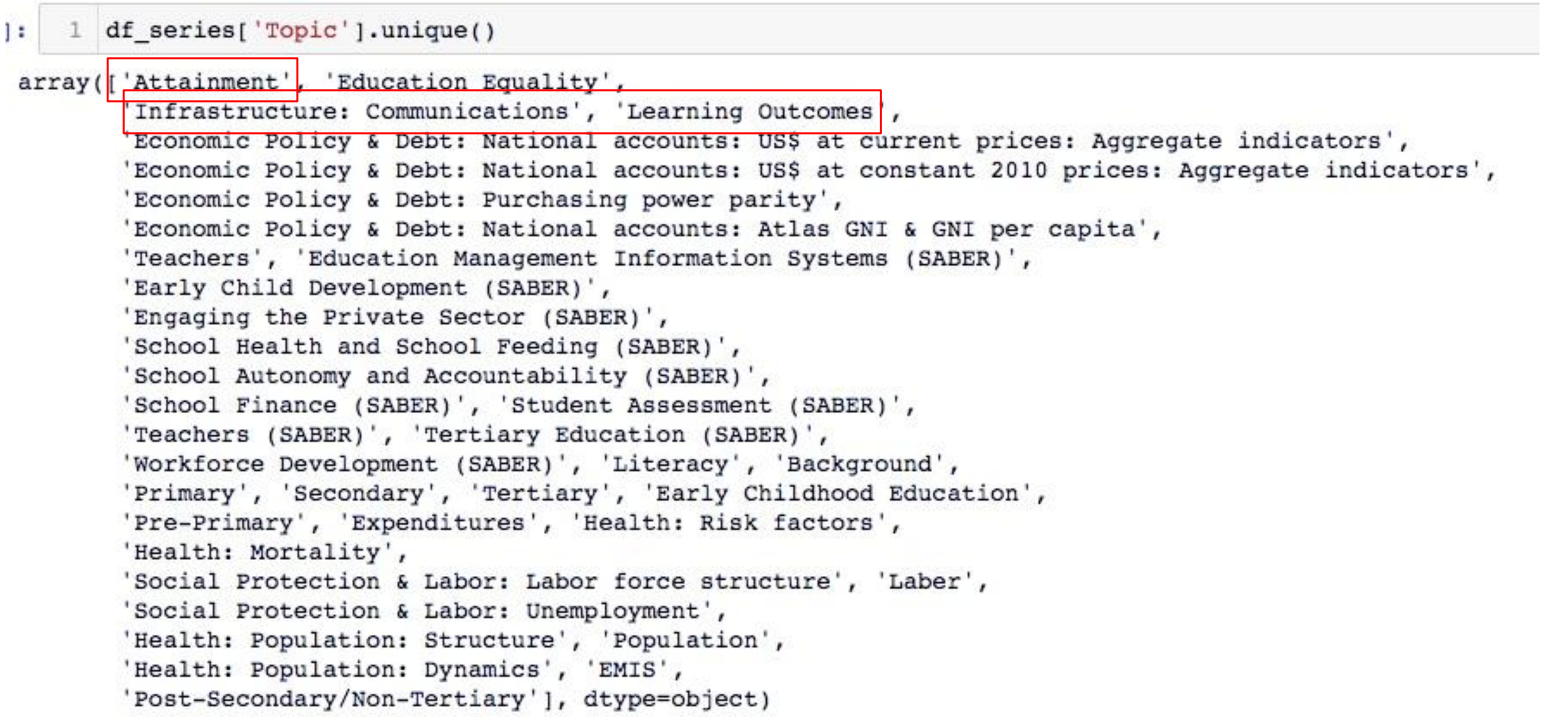
```
df_series['Topic'].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



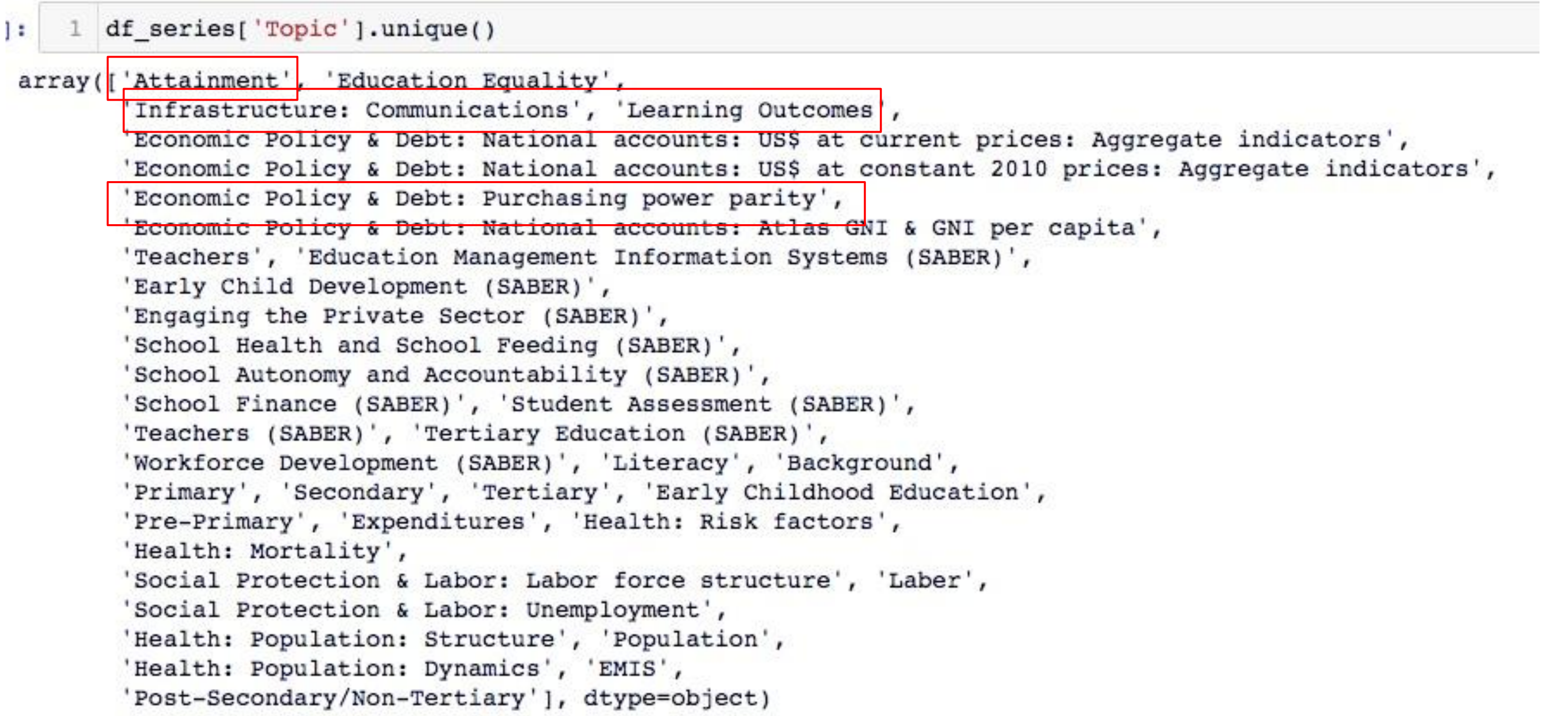
```
]:  df_series['Topic'].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



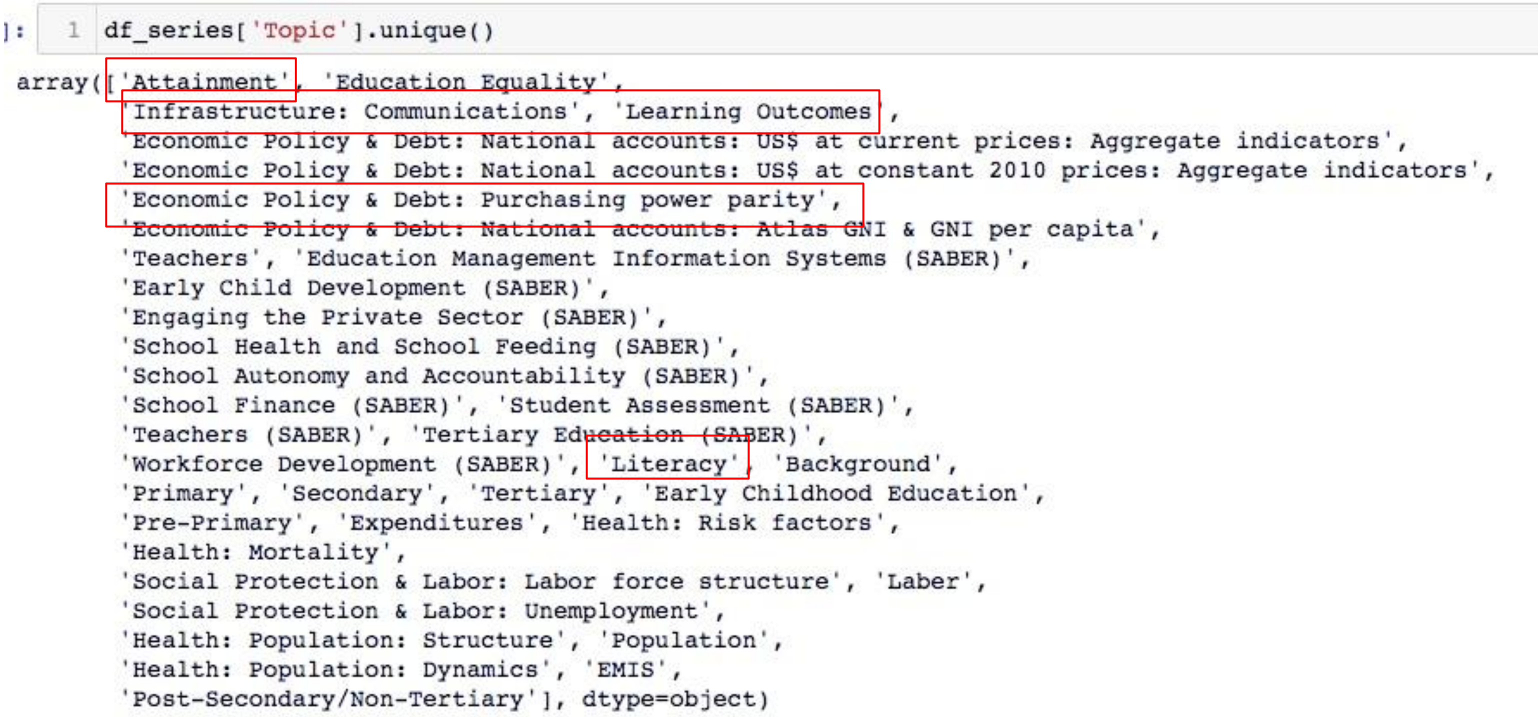
```
df_series['Topic'].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



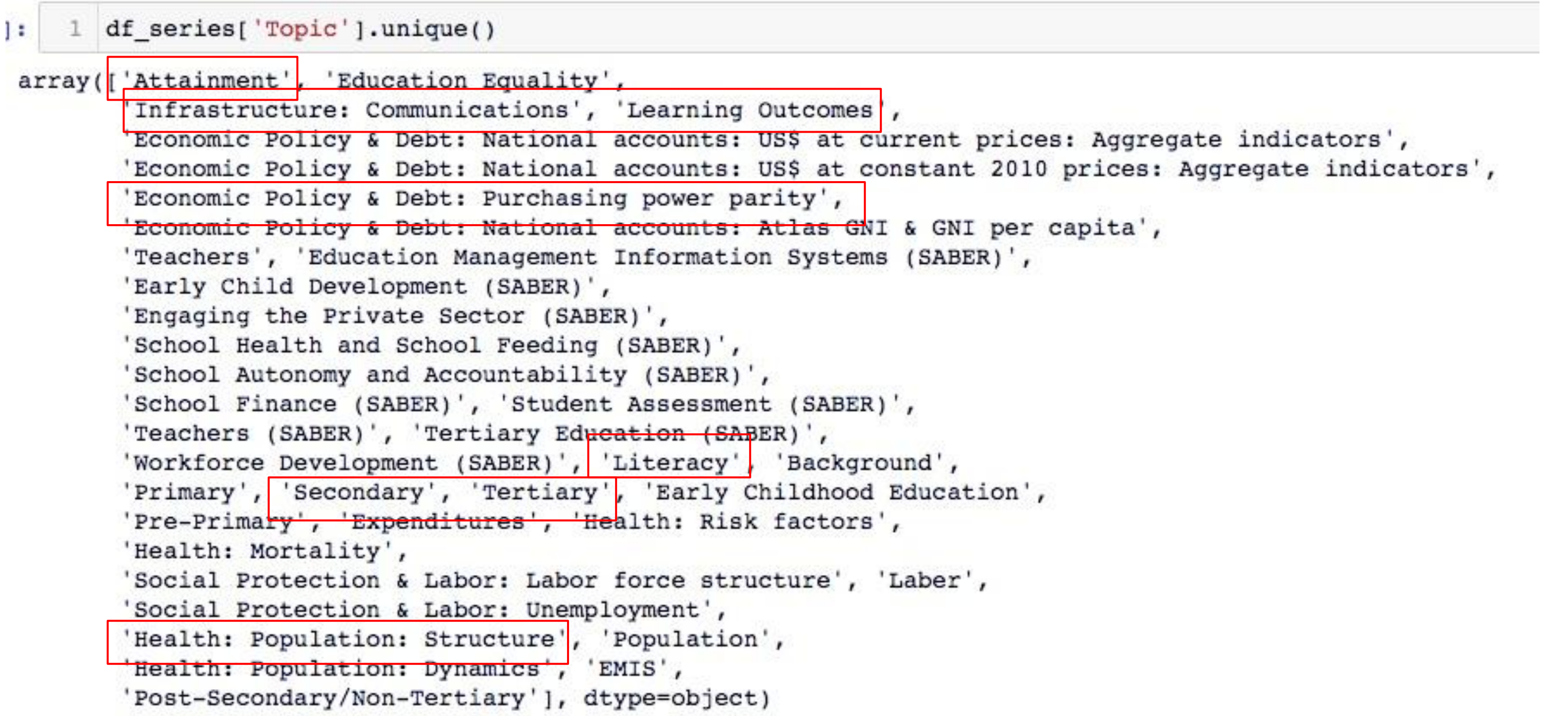
```
]:  df_series['Topic'].unique()
```

```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs: choix des topics



```
]:  df_series['Topic'].unique()
```

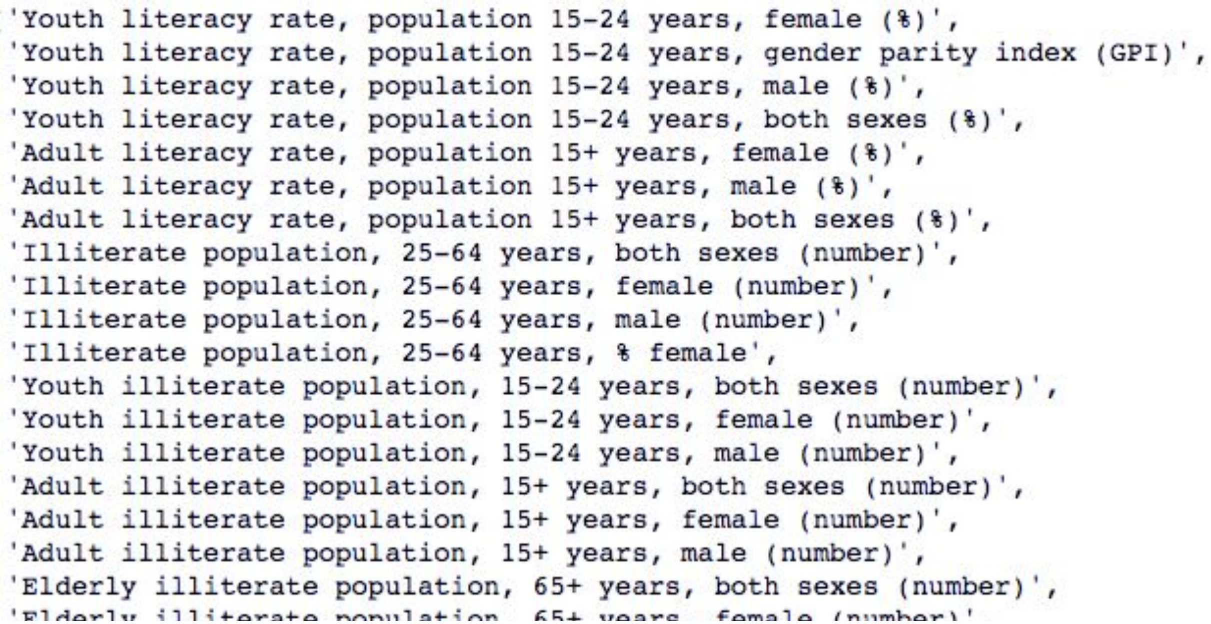
```
array(['Attainment', 'Education Equality',
       'Infrastructure: Communications', 'Learning Outcomes',
       'Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators',
       'Economic Policy & Debt: National accounts: US$ at constant 2010 prices: Aggregate indicators',
       'Economic Policy & Debt: Purchasing power parity',
       'Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita',
       'Teachers', 'Education Management Information Systems (SABER)',
       'Early Child Development (SABER)',
       'Engaging the Private Sector (SABER)',
       'School Health and School Feeding (SABER)',
       'School Autonomy and Accountability (SABER)',
       'School Finance (SABER)', 'Student Assessment (SABER)',
       'Teachers (SABER)', 'Tertiary Education (SABER)',
       'Workforce Development (SABER)', 'Literacy', 'Background',
       'Primary', 'Secondary', 'Tertiary', 'Early Childhood Education',
       'Pre-Primary', 'Expenditures', 'Health: Risk factors',
       'Health: Mortality',
       'Social Protection & Labor: Labor force structure', 'Laber',
       'Social Protection & Labor: Unemployment',
       'Health: Population: Structure', 'Population',
       'Health: Population: Dynamics', 'EMIS',
       'Post-Secondary/Non-Tertiary'], dtype=object)
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs

Exemple: topic “Literacy”



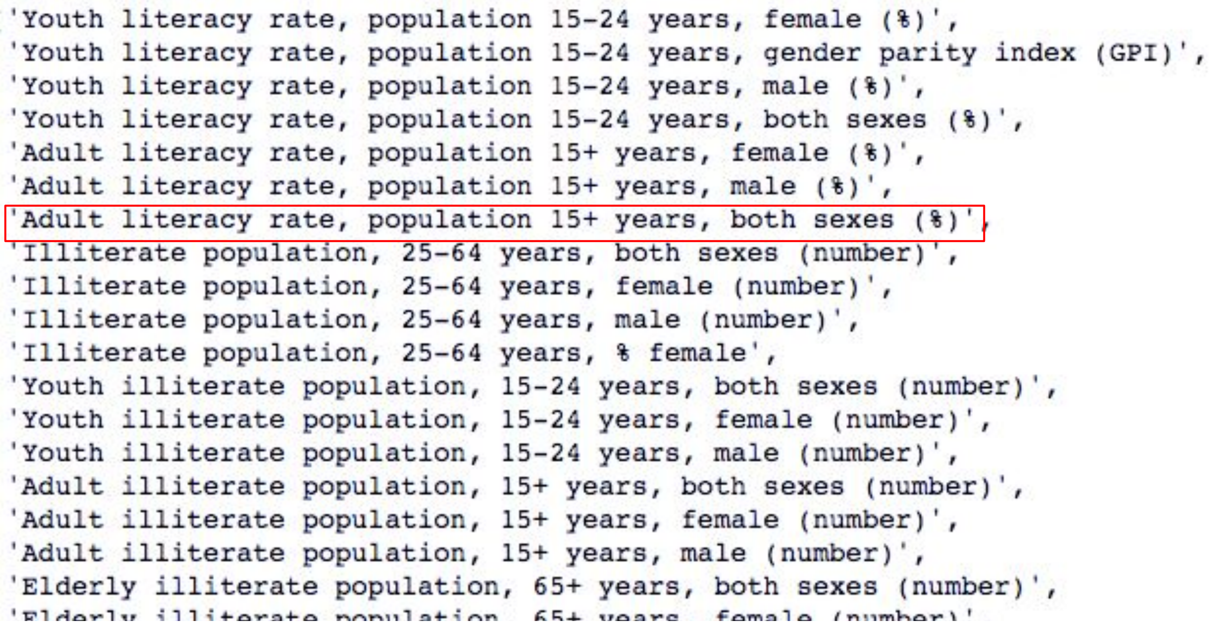
```
'Youth literacy rate, population 15-24 years, female (%)',
'Youth literacy rate, population 15-24 years, gender parity index (GPI)',
'Youth literacy rate, population 15-24 years, male (%)',
'Youth literacy rate, population 15-24 years, both sexes (%)',
'Adult literacy rate, population 15+ years, female (%)',
'Adult literacy rate, population 15+ years, male (%)',
'Adult literacy rate, population 15+ years, both sexes (%)',
'Illiterate population, 25-64 years, both sexes (number)',
'Illiterate population, 25-64 years, female (number)',
'Illiterate population, 25-64 years, male (number)',
'Illiterate population, 25-64 years, % female',
'Youth illiterate population, 15-24 years, both sexes (number)',
'Youth illiterate population, 15-24 years, female (number)',
'Youth illiterate population, 15-24 years, male (number)',
'Adult illiterate population, 15+ years, both sexes (number)',
'Adult illiterate population, 15+ years, female (number)',
'Adult illiterate population, 15+ years, male (number)',
'Elderly illiterate population, 65+ years, both sexes (number)',
'Elderly illiterate population, 65+ years, female (number)'
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs

Exemple: topic “Literacy”



```
'Youth literacy rate, population 15-24 years, female (%)',
'Youth literacy rate, population 15-24 years, gender parity index (GPI)',
'Youth literacy rate, population 15-24 years, male (%)',
'Youth literacy rate, population 15-24 years, both sexes (%)',
'Adult literacy rate, population 15+ years, female (%)',
'Adult literacy rate, population 15+ years, male (%)',
'Adult literacy rate, population 15+ years, both sexes (%)',
'Illiterate population, 25-64 years, both sexes (number)',
'Illiterate population, 25-64 years, female (number)',
'Illiterate population, 25-64 years, male (number)',
'Illiterate population, 25-64 years, % female',
'Youth illiterate population, 15-24 years, both sexes (number)',
'Youth illiterate population, 15-24 years, female (number)',
'Youth illiterate population, 15-24 years, male (number)',
'Adult illiterate population, 15+ years, both sexes (number)',
'Adult illiterate population, 15+ years, female (number)',
'Adult illiterate population, 15+ years, male (number)',
'Elderly illiterate population, 65+ years, both sexes (number)',
'Elderly illiterate population, 65+ years, female (number)'
```

Analyse pré-exploratoire
Nettoyage du jeu de données

Sélection des indicateurs

- ➢ Enrolment in secondary education, both sexes (number)
- ➢ Enrolment in lower secondary education, both sexes (number)
- ➢ Enrolment in upper secondary education, both sexes (number)
- ➢ Enrolment in tertiary education, all programs, both sexes (number)
- ➢ GDP per capita, PPP (constant 2011 international $)
- ➢ Expenditure on tertiary as % of government expenditure on education (%)
- ➢ Government expenditure per upper secondary student (PPP$)
- ➢ Government expenditure per tertiary student (PPP$)
- ➢ Personal computers (per 100 people)
- ➢ Internet users (per 100 people)
- ➢ Adult literacy rate, population 15+ years, both sexes (%)
- ➢ Population of the official age for lower secondary education, both sexes (number)
- ➢ Population of the official age for upper secondary education, both sexes (number)
- ➢ Population of the official age for tertiary education, both sexes (number)

Analyse pré-exploratoire

Nettoyage du jeu de données

- Filtrage pays/régions
- Restriction du jeu de données aux indicateurs présélectionnés
- Restriction du jeu de données aux données passées, et non prospectives
    - données chaque année entre 1970 et 2017, puis 2020 puis tous les 5 ans jusqu’en 2100
    - données prospectives absentes pour les indicateurs sélectionnés
- Restriction aux années 2000-2016:
    - Données trop anciennes peu utiles pour notre étude
    - Données mieux renseignées à partir de 2000
    - Données absentes pour 2017

Point de départ 2000 pour garantir qu’on dispose de données:

- données relativement récentes
- données effectivement renseignées, au moins pour une année entre 2000 et 2016

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par la population

- Filtre sur l’indicateur “Population of the official age for upper secondary education, both sexes (number) “
- 1er décile: 11572 personnes

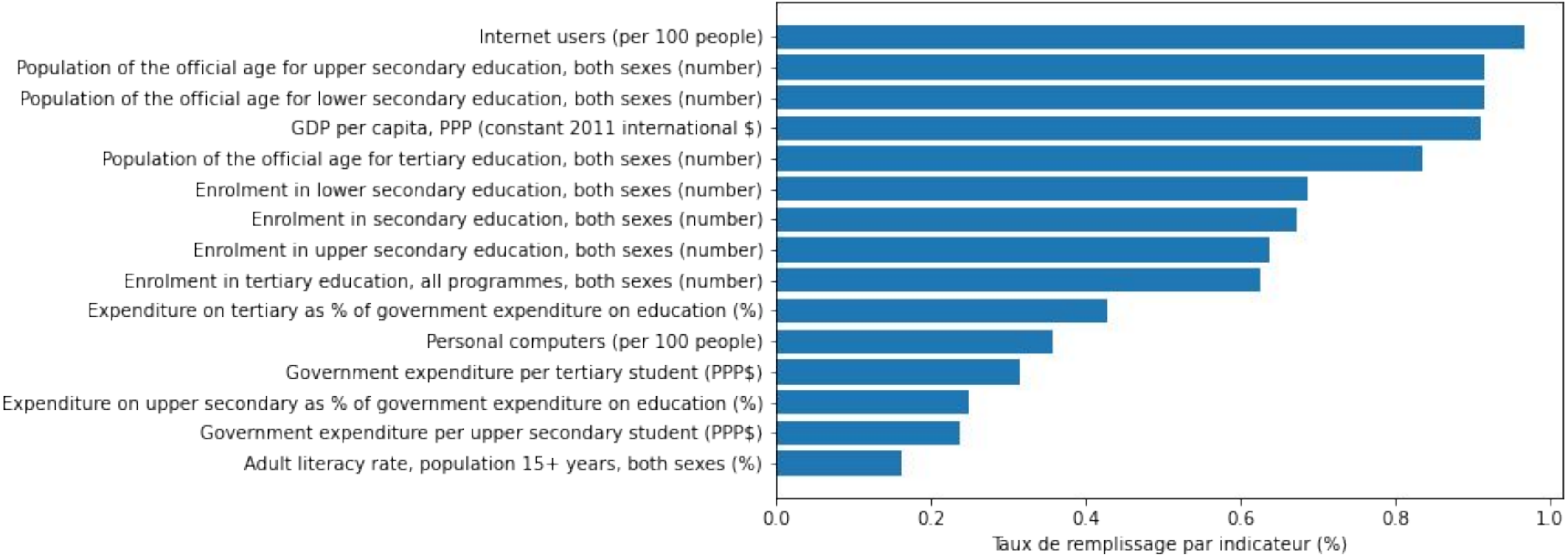
➢ On élimine les pays appartenant au premier décile:
  - Aruba
  - Antigua and Barbuda
  - Belize
  - Bermuda
  - Barbados
  - Curacao
  - Dominica
  - Micronesia, Fed. Sts.
  - Gibraltar
  - Grenada
  - Kiribati
  - St. Kitts and Nevis
  - St. Lucia
  - Liechtenstein
  - Marshall Islands
  - Malta
  - Palau
  - San Marino
  - Sao Tome and Principe
  - Seychelles
  - Tonga
  - Tuvalu
  - St. Vincent and the Grenadines

- Soit environ 10% des pays renseignés dans le jeu de données

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage: Taux de remplissage par indicateur

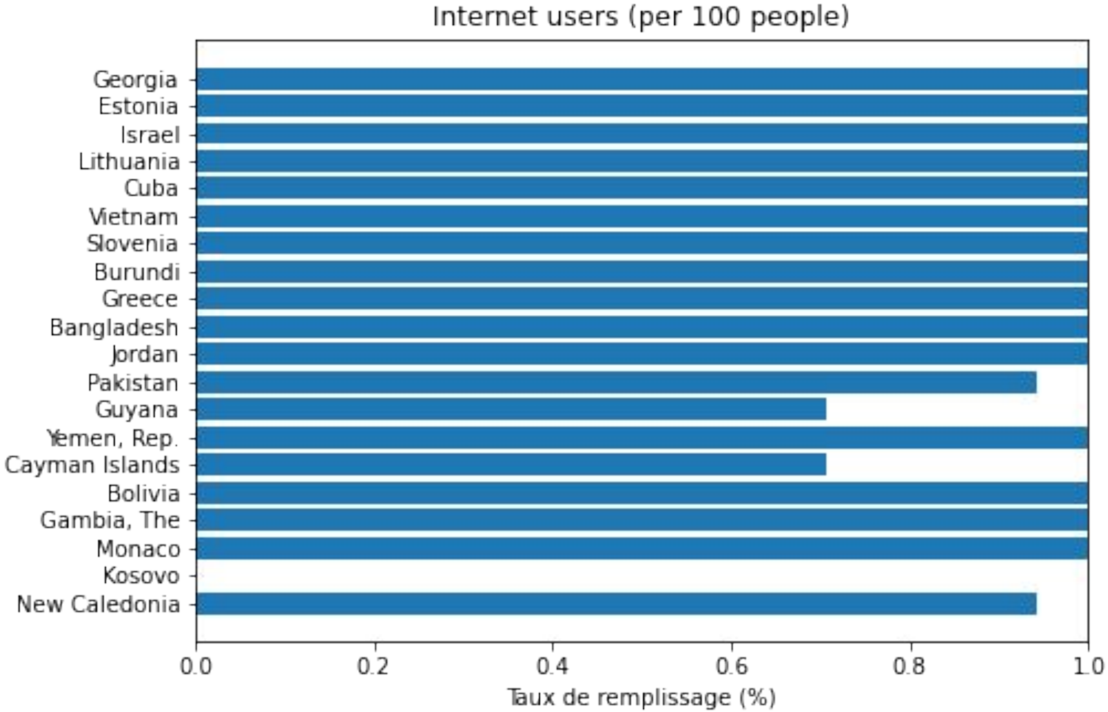


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage: Taux de remplissage par indicateur et pays

Exemple sur 3 indicateurs:

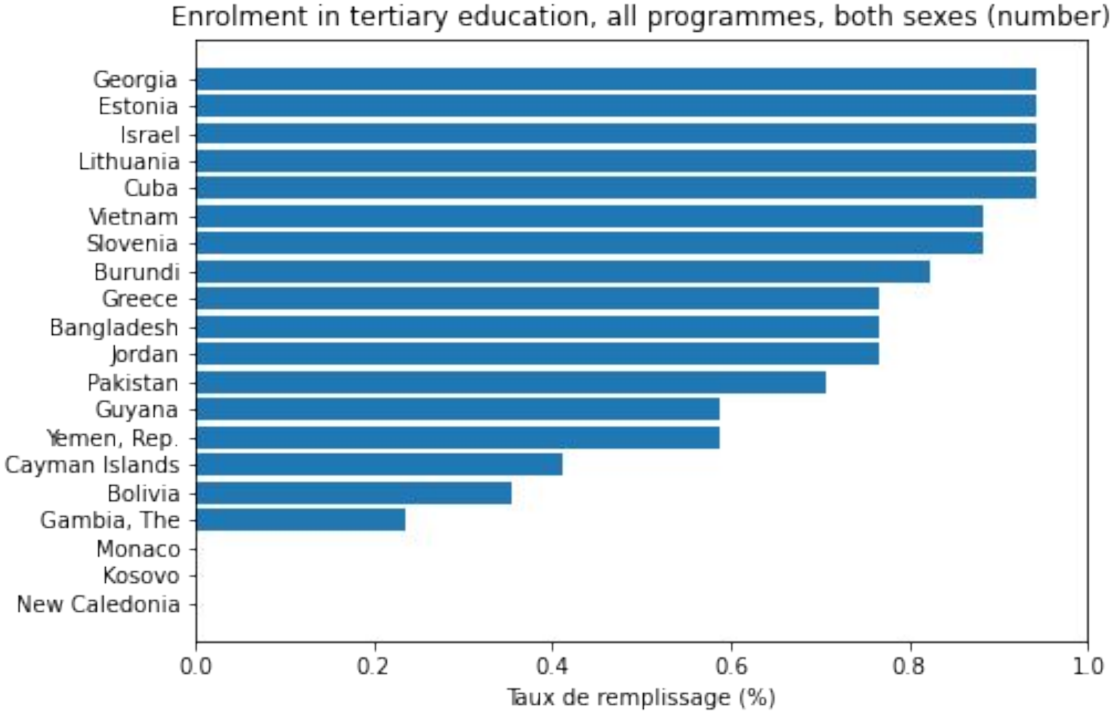


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage: Taux de remplissage par indicateur et pays

Exemple sur 3 indicateurs:

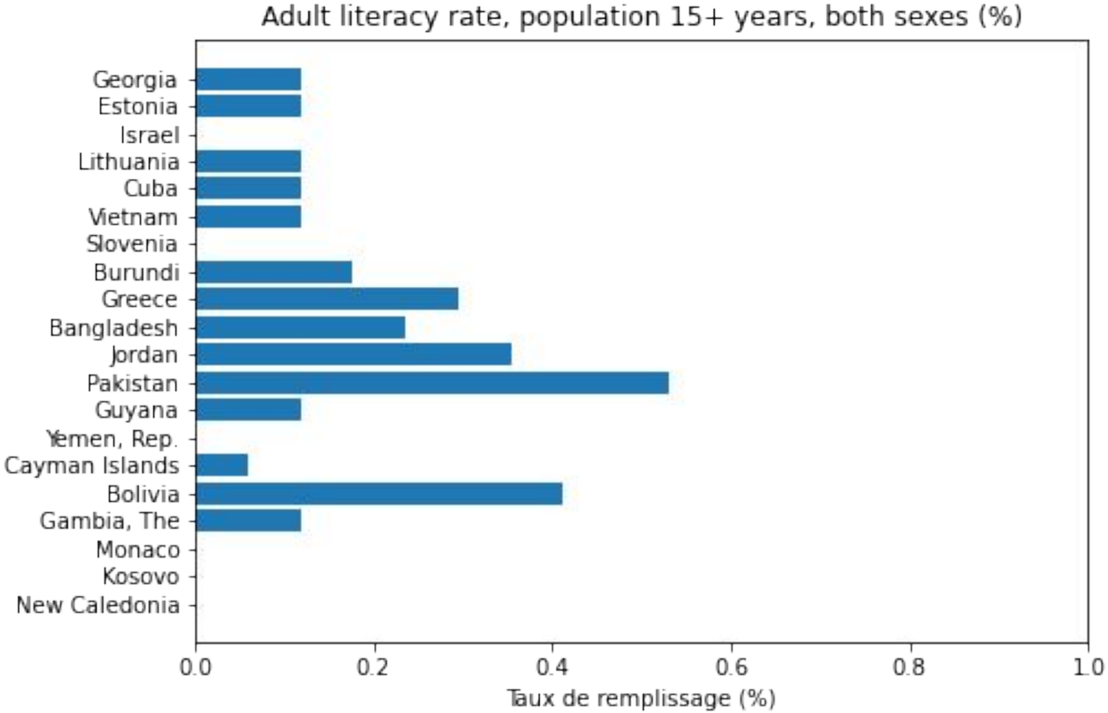


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage: Taux de remplissage par indicateur et pays

Exemple sur 3 indicateurs:



Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

Analyse pré-exploratoire

Nettoyage du jeu de données

Statistiques pays

Analyse pré-exploratoire

Nettoyage du jeu de données

Statistiques régions

Analyse pré-exploratoire

Nettoyage du jeu de données

Scoring

Analyse pré-exploratoire

Nettoyage du jeu de données

Proposition de pays avec du potentiel

Questions

Analyse pré-exploratoire

Nettoyage du jeu de données

Années I: données passées vs. données prospectives

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection du dataset pays/données passées

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection du dataset pays/données passées

Analyse pré-exploratoire

Nettoyage du jeu de données

Années 2: sélection des années 2000-2016