Analyse de données de systèmes éducatifs pour academy

Olivier Legrand

OPENCLASSROOMS

Sommaire

1. Présentation du jeu de données
2. Présentation de l’analyse pré-exploratoire
3. Pertinence du jeu de données
4. Questions

Présentation du jeu de données

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

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	Nal
			Adjusted															

- 70 colonnes: 4 colonnes Country Name & Code, Indicator Name & Code, puis années 1970 → 2100
- 886930 lignes
- Beaucoup de données manquantes, pas de doublons
- 3665 Indicator Name uniques

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..

- 21 colonnes, seulement 6 remplies à plus de 15%:
  - Series code, Topic, Indicator Name, Short & Long definition, Source

- 3665 lignes - les “Indicators” rencontrés précédemment

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

- renseigne sur les pays présents dans le dataset EdStatsData
- 241 lignes: pays du monde et régions/continent/entités géopolitiques (comme l'UE par exemple).
- Colonnes totalement remplies: Country Code, Short Name, Table Name, Long Name
- Colonnes intéressantes a priori:
  - Currency Unit
  - Region
  - Income Group
  - Latest population census
  - Source of most recent Income and expenditure data

Présentation du jeu de données

- EdStatscountrySeries & EdStatsFootNote

- Tables qui renseignent sur certains indicateurs (source, description)
- peu utiles dans le cadre de cette analyse

Analyse pré-exploratoire

- Nettoyage du dataset
  - Indicateurs et thèmes (topics)
  - Filtrage pays/régions
  - Restriction aux indicateurs sélectionnés
  - Restriction aux données historiques (non prospectives)
  - Restriction aux 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

```
1]: 1 print(df_series['Topic'].unique())
```

```
['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']
```

Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection des indicateurs. Exemple: topic “Literacy”

```
outh 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)'
'Elderly illiterate population, 65+ years, male (number)'
'Youth illiterate population, 15-24 years, % female'
lt illiterate population, 15+ years, % female'
```

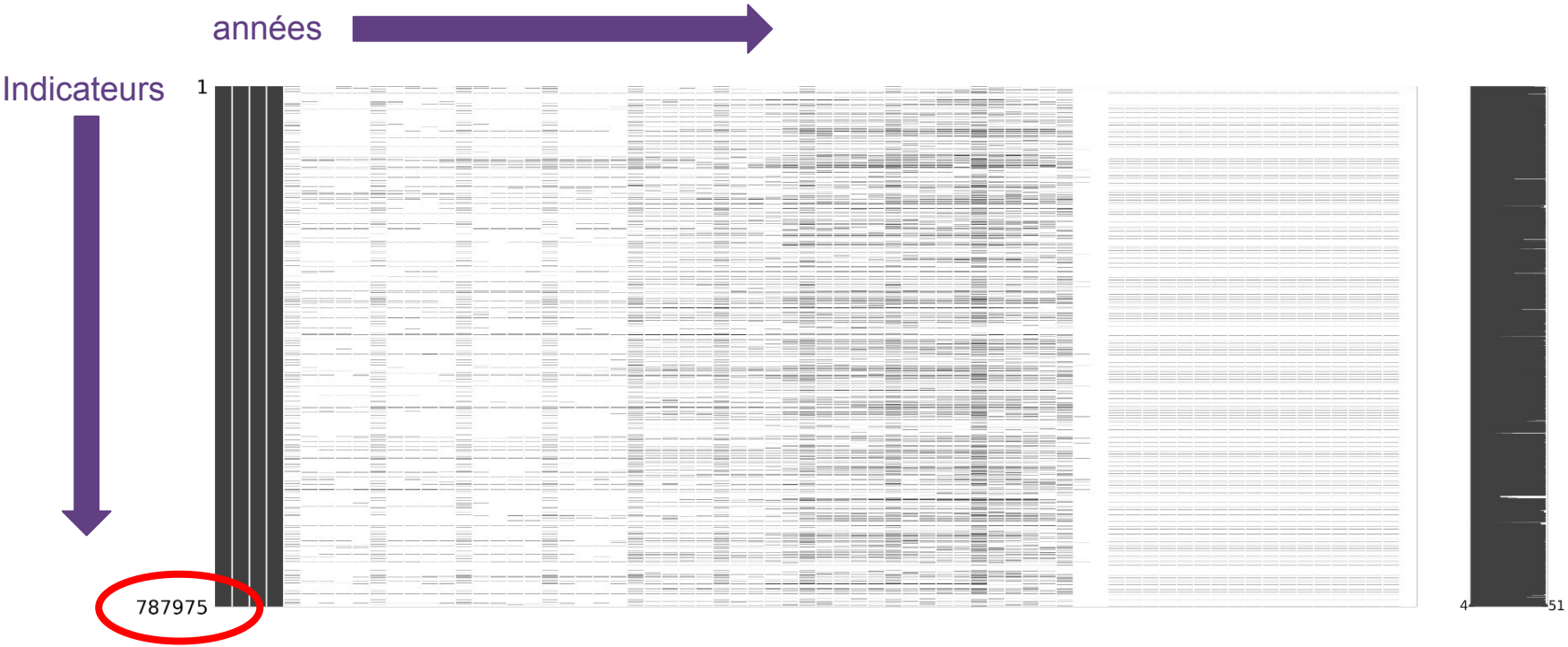
Analyse pré-exploratoire
Nettoyage du jeu de données
Sélection des indicateurs

- 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 $)
- Expenditure on tertiary as % of government expenditure on education (%)
- Expenditure on upper secondary as % of government expenditure on education (%)
- Government expenditure per upper secondary student (PPP$)
- 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

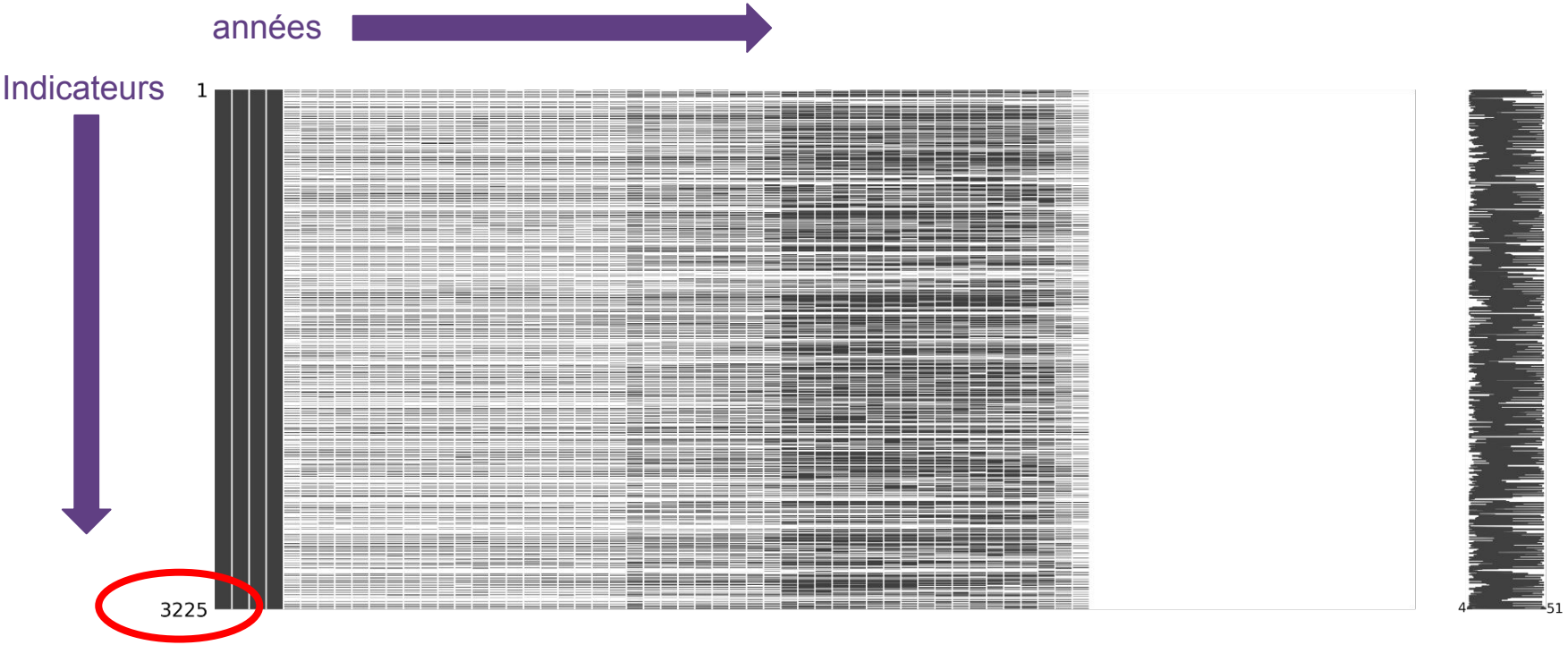
➢ Filtrage pays/régions: sélection des pays uniquement



Analyse pré-exploratoire

Nettoyage du jeu de données

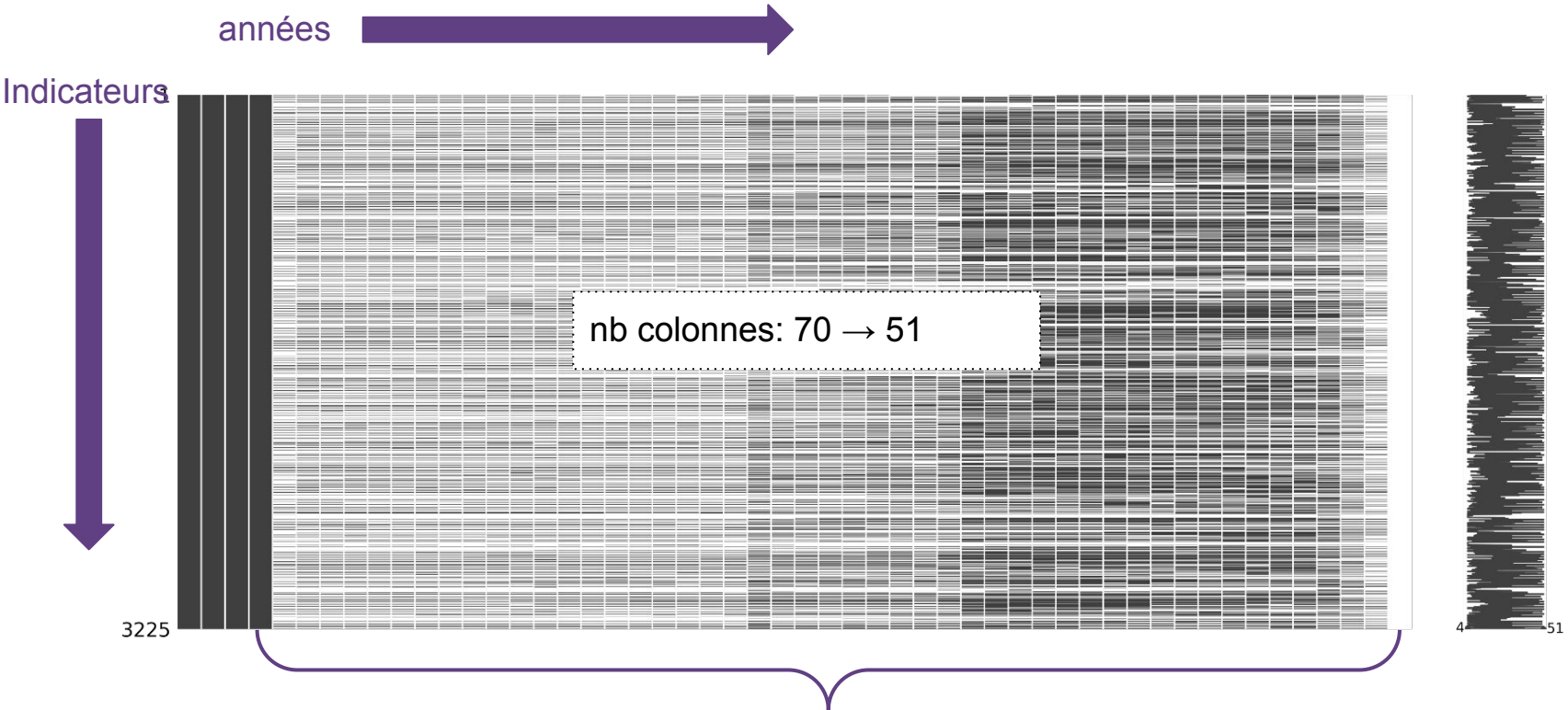
➢ Restriction du jeu de données aux indicateurs présélectionnés



Analyse pré-exploratoire

Nettoyage du jeu de données

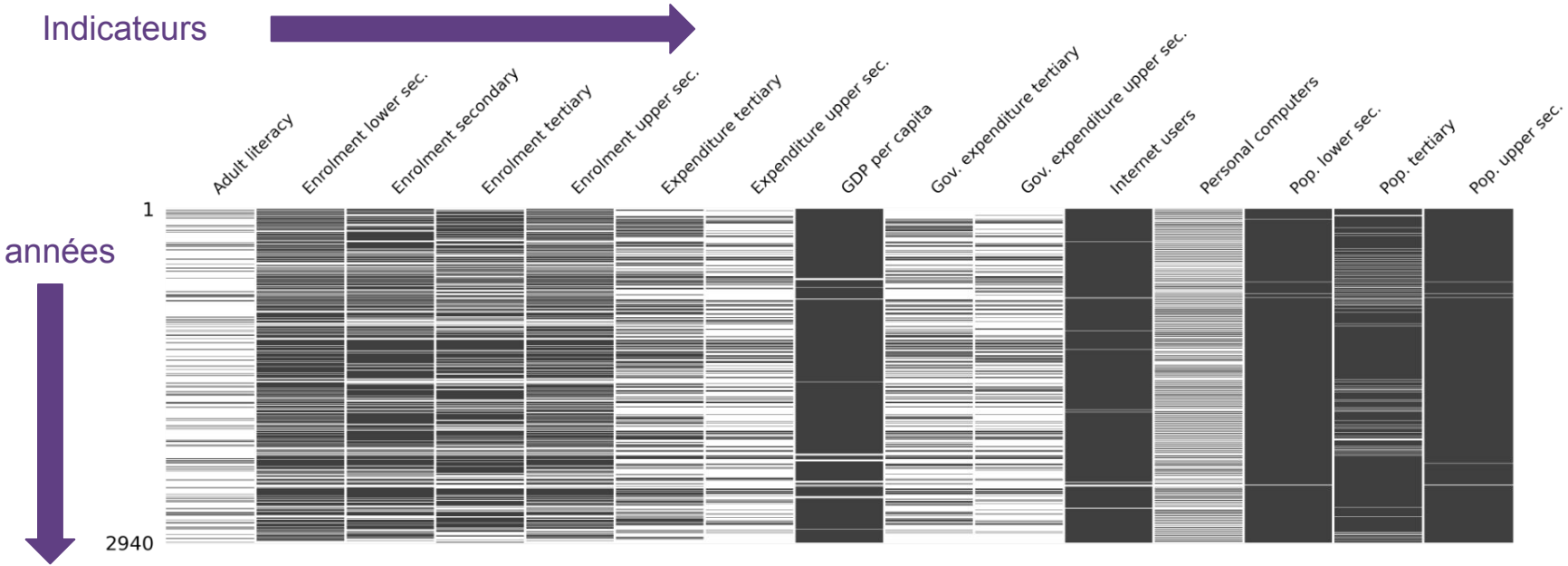
➢ Restriction du jeu de données aux données historiques, et non prospectives



Analyse pré-exploratoire

Nettoyage du jeu de données

- Restriction aux années 2000-2016
- Restriction aux pays qui possèdent des données de population entre 2000 et 2016, et qui ne sont pas trop petits (pop > 11572 hab. i.e. 1er décile):

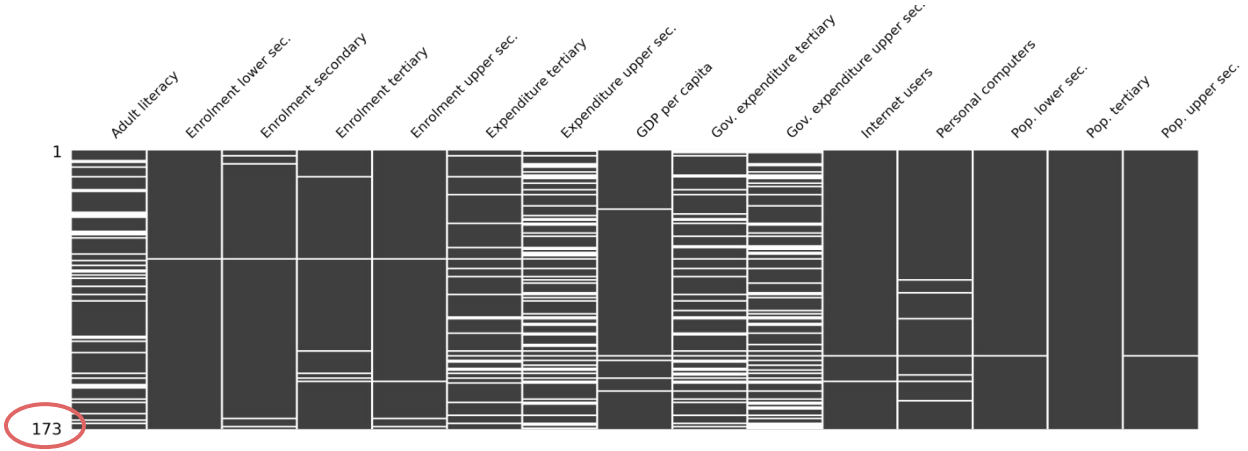


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par le taux de remplissage

- Pour chaque pays et chaque indicateur, on conserve l'année la plus récente **et** la mieux remplie:
  - pour chaque pays, dernière année renseignée (pour chaque indicateur)
  - pour chaque indicateur, dernières années renseignées dans [2000, 2016]
  - pour chaque indicateur, année de référence = "dernière année" la plus souvent citée

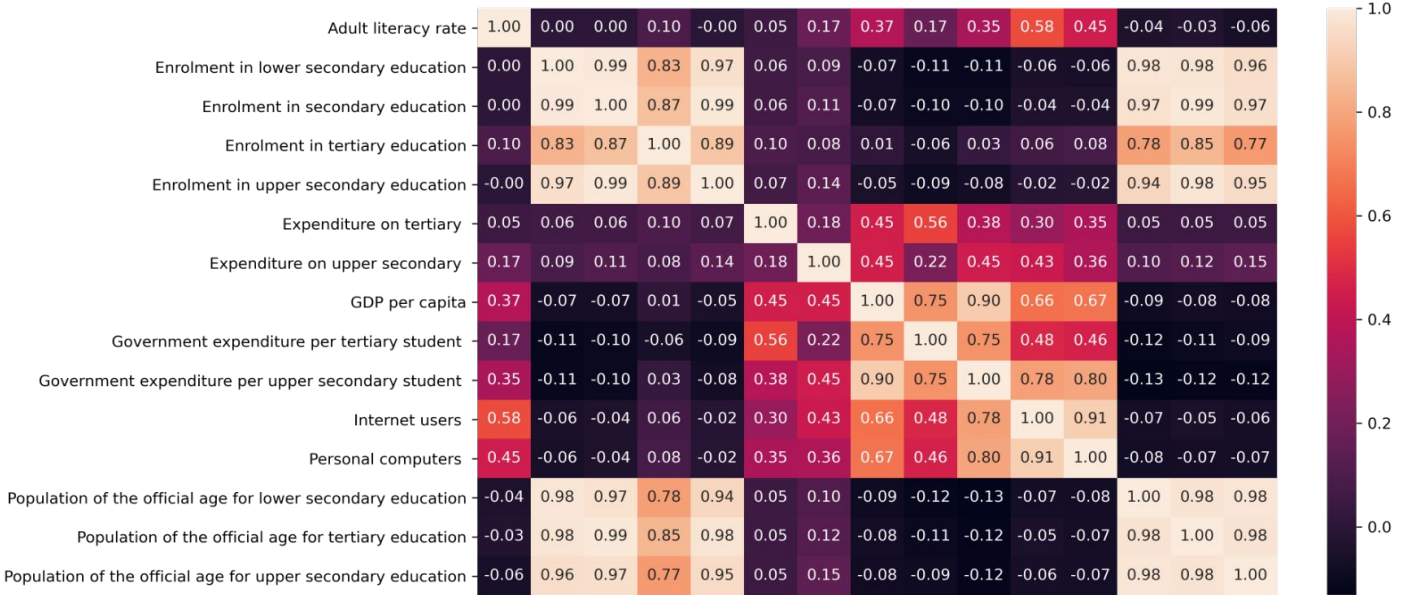


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

- Pour chaque topic, sélection d'un représentant:
  - fortement corrélé avec les autres indicateurs
  - bien renseigné
  - idéalement, peu corrélé avec les autres topics

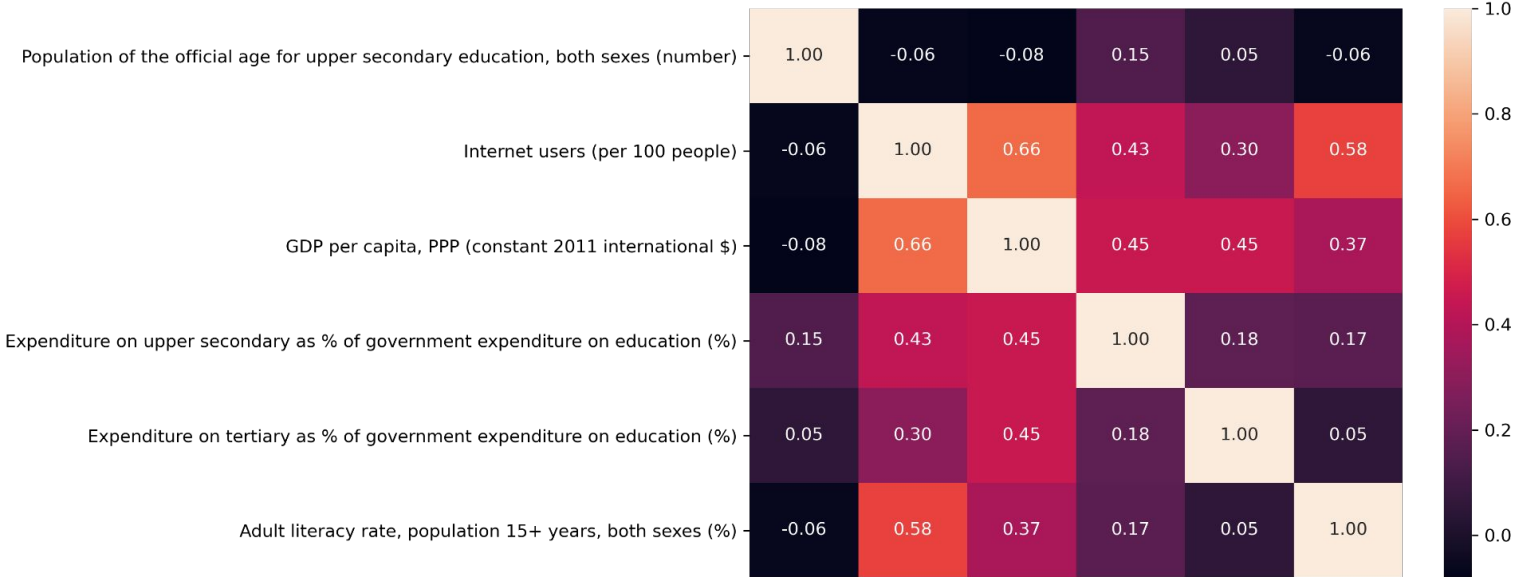


Analyse pré-exploratoire

Nettoyage du jeu de données

Sélection par les corrélations

- Pour chaque topic, sélection d'un représentant:
  - fortement corrélé avec les autres indicateurs
  - bien renseigné
  - idéalement, peu corrélé avec les autres topics

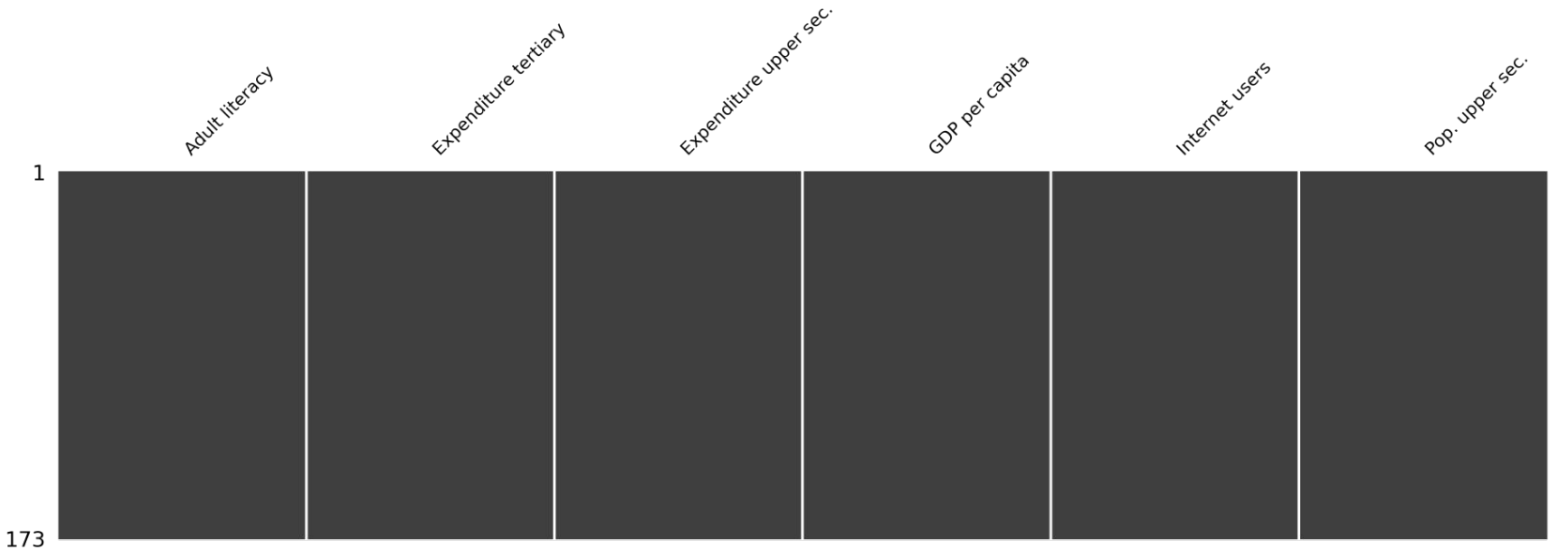


Analyse pré-exploratoire

Nettoyage du jeu de données

Interpolation des valeurs manquantes

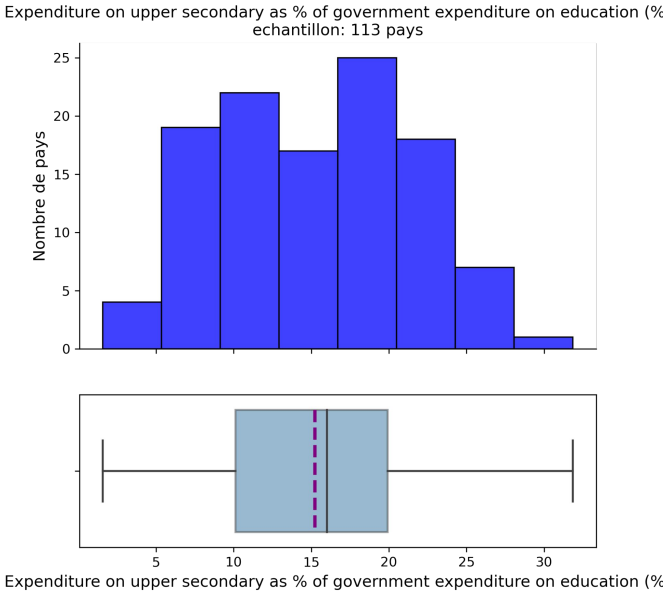
- Interpolation réalisée grâce aux valeurs moyennes de chaque région/income group pour chaque indicateur



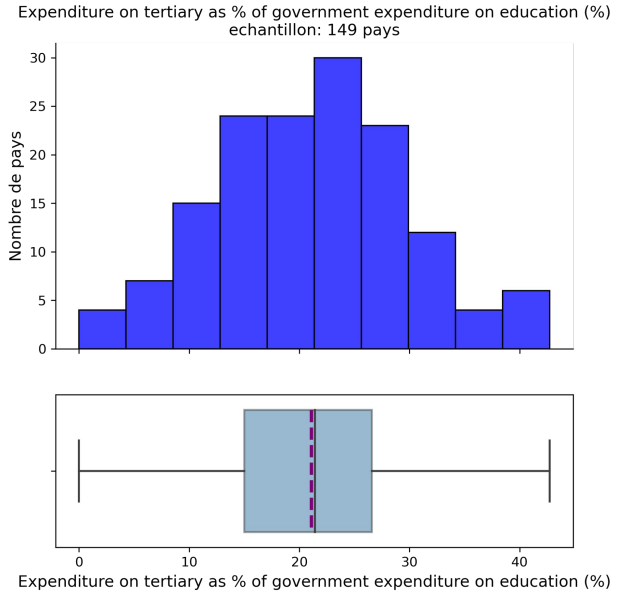
Pertinence du jeu de données: des statistiques en lien avec l'éducation - exemples d'indicateurs

Statistiques Pays

Dépenses pour le lycée en % des dépenses publiques pour l'éducation



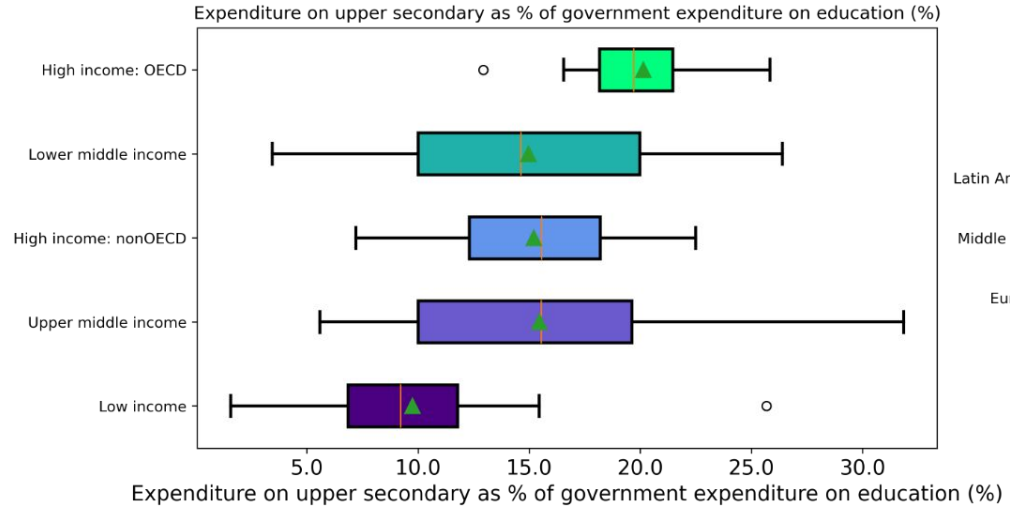
Dépenses pour le supérieur en % des dépenses publiques pour l'éducation



Pertinence du jeu de données: des statistiques au niveau régional, ou par agrégat

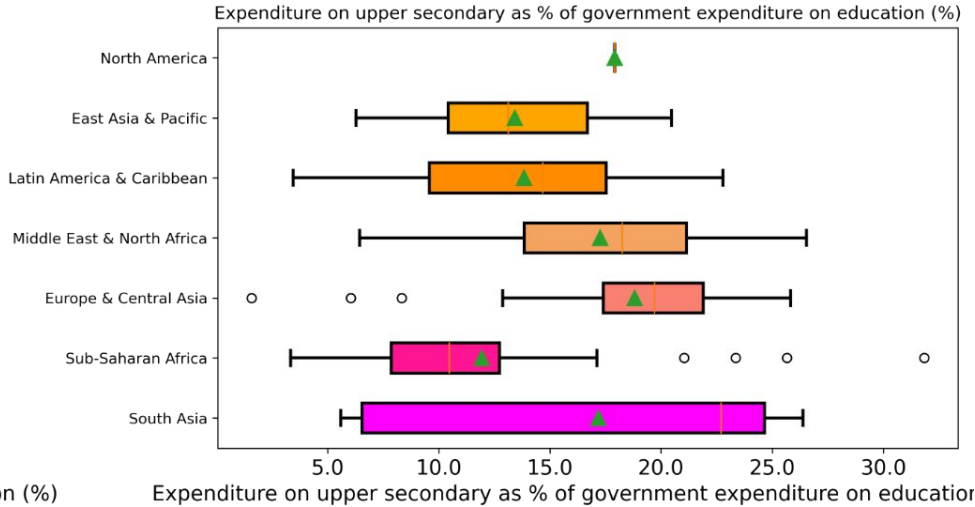
Statistiques agrégat économique

Dépenses pour le lycée en % des dépenses publiques pour l'éducation



Statistiques Régions

Dépenses pour le lycée en % des dépenses publiques pour l'éducation

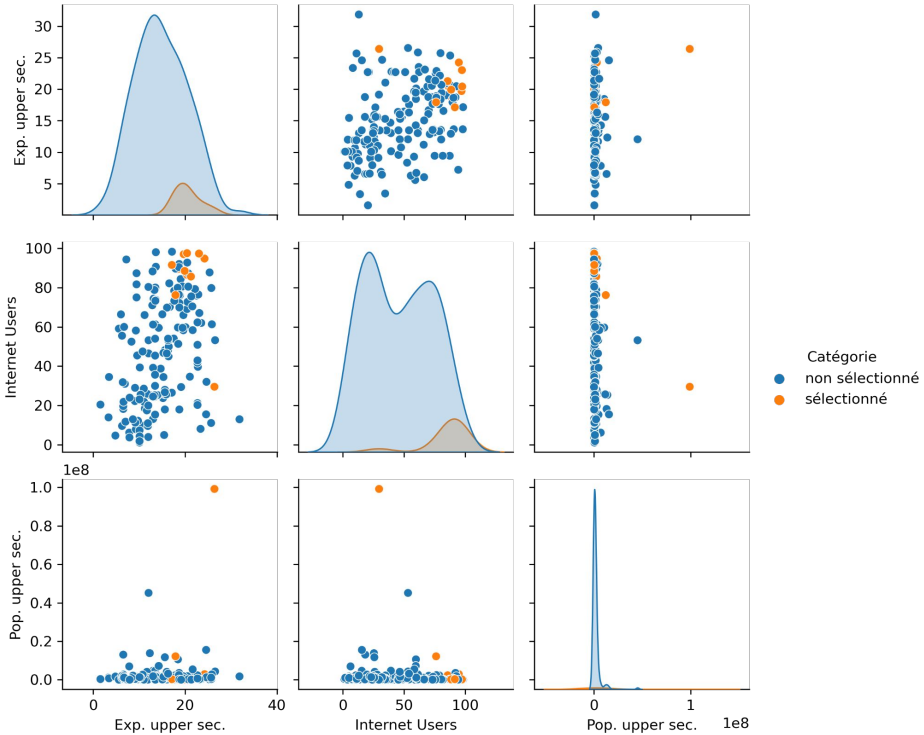


Pertinence du jeu de données - Classement

- Renormalisation des valeurs pour chaque indicateur: X-min/Max-Min
- Pour chaque pays, somme sur toutes les valeurs
- Critère de sélection: meilleur score entre 0 et 6 et langue parlée français ou anglais
- Résultat:
  - Luxembourg
  - Canada
  - United States
  - United Kingdom
  - Switzerland
  - India
  - France
  - New Zealand
  - Belgium

Pertinence du jeu de données - Classement

Visualisation des résultats sur 3 indicateurs



Pertinence du jeu de données - Conclusion

- Plus:
  - Jeu de données important: 3665 indicateurs, 241 pays/régions, 65 années
  - Beaucoup de nettoyage pour pouvoir l’exploiter
  - indicateurs variés et pertinents
  - permet d’obtenir des infos au niveau du pays, de la région et sur des agrégats économiques
  - permet d’effectuer un classement préliminaire
- Moins:
  - Beaucoup de nettoyage pour pouvoir l’exploiter
  - pas d’indicateur “business” ou “langue” inclus

Questions

Pertinence du jeu de données - Potentiel de croissance

Potentiel de croissance des pays sélectionnés - données de la période 2000-2016

Pertinence du jeu de données - Potentiel de croissance

Potentiel de croissance des pays sélectionnés - données de la période 2000-2016

Analyse pré-exploratoire

Nettoyage du jeu de données

Interpolation des valeurs manquantes

- Interpolation réalisée grâce aux valeurs moyennes de chaque région/income group pour chaque indicateur

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