2019

OPENCLASSROOMS

Parcours Data-Scientist – Projet 2

ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

PROBLÉMATIQUE

SOMMAIRE

Mission d’analyse exploratoire à partir des données de la banque mondiale :

- *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 ?*

Présentation des datasets :

- Structure
- Contenus

Structure des datasets

Pandas.info() :

EdStatsCountry.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241 entries, 0 to 240
Data columns (total 32 columns):
Country Code                                     241 non-null object
Short Name                                       241 non-null object
Table Name                                       241 non-null object
Long Name                                        241 non-null object
2-alpha code                                     238 non-null object
Currency Unit                                    215 non-null object
Special Notes                                    145 non-null object
Region                                           214 non-null object
Income Group                                     214 non-null object
WB-2 code                                        240 non-null object
National accounts base year                      205 non-null object
National accounts reference year                 32 non-null float64
SNA price valuation                              197 non-null object
Lending category                                 144 non-null object
Other groups                                     58 non-null object
System of National Accounts                      215 non-null object
Alternative conversion factor                    47 non-null object
PPP survey year                                  145 non-null object
Balance of Payments Manual in use                181 non-null object
External debt Reporting status                   124 non-null object
System of trade                                  200 non-null object
Government Accounting concept                    161 non-null object
IMF data dissemination standard                  181 non-null object
Latest population census                         213 non-null object
Latest household survey                          141 non-null object
Source of most recent Income and expenditure data    160 non-null object
Vital registration complete                      111 non-null object
Latest agricultural census                       142 non-null object
Latest industrial data                           107 non-null float64
Latest trade data                                185 non-null float64
Latest water withdrawal data                     179 non-null object
Unnamed: 31                                      0 non-null float64
dtypes: float64(4), object(28)
memory usage: 60.4+ KB
```

EdStatsCountry-Series.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 613 entries, 0 to 612
Data columns (total 4 columns):
CountryCode    613 non-null object
SeriesCode     613 non-null object
DESCRIPTION    613 non-null object
Unnamed: 3     0 non-null float64
dtypes: float64(1), object(3)
memory usage: 19.3+ KB
```

EdStatsData.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 70 columns):
Country Name      886930 non-null object
Country Code      886930 non-null object
Indicator Name    886930 non-null object
Indicator Code    886930 non-null object
1970              72288 non-null float64
1971              35537 non-null float64
1972              35619 non-null float64
1973              35545 non-null float64
1974              35730 non-null float64
1975              87306 non-null float64
1976              37483 non-null float64
1977              37574 non-null float64
```

EdStatsSeries.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Data columns (total 21 columns):
Series Code                               3665 non-null object
Topic                                     3665 non-null object
Indicator Name                            3665 non-null object
Short definition                          2156 non-null object
Long definition                           3665 non-null object
Unit of measure                           0 non-null float64
Periodicity                               99 non-null object
Base Period                               314 non-null object
Other notes                               552 non-null object
Aggregation method                        47 non-null object
Limitations and exceptions                14 non-null object
Notes from original source                0 non-null float64
General comments                          14 non-null object
Source                                    3665 non-null object
Statistical concept and methodology       23 non-null object
Development relevance                     3 non-null object
Related source links                      215 non-null object
Other web links                           0 non-null float64
Related indicators                        0 non-null float64
License Type                              0 non-null float64
Unnamed: 20                               0 non-null float64
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```

EdStatsFootNote.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 643638 entries, 0 to 643637
Data columns (total 5 columns):
CountryCode    643638 non-null object
SeriesCode     643638 non-null object
Year           643638 non-null object
DESCRIPTION    643638 non-null object
Unnamed: 4     0 non-null float64
dtypes: float64(1), object(4)
memory usage: 24.6+ MB
```

PRÉSENTATION DES DONNÉES

Contenu des datasets

Intéressant
Intéressant mais redondant/manquant

Pandas.info() :

EdStatsCountry.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241 entries, 0 to 240
Data columns (total 32 columns):
Country Code                                     241 non-null object
Short Name                                       241 non-null object
Table Name                                       241 non-null object
Long Name                                        241 non-null object
2-alpha code                                     238 non-null object
Currency Unit                                    215 non-null object
Special Notes                                    145 non-null object
Region                                           214 non-null object
Income Group                                     214 non-null object
WB-2 code                                        240 non-null object
National accounts base year                      205 non-null object
National accounts reference year                 32 non-null float64
SNA price valuation                              197 non-null object
Lending category                                 144 non-null object
Other groups                                     58 non-null object
System of National Accounts                      215 non-null object
Alternative conversion factor                    47 non-null object
PPP survey year                                  145 non-null object
Balance of Payments Manual in use                181 non-null object
External debt Reporting status                   124 non-null object
System of trade                                  200 non-null object
Government Accounting concept                    161 non-null object
IMF data dissemination standard                  181 non-null object
Latest population census                         213 non-null object
Latest household survey                          141 non-null object
Source of most recent Income and expenditure data 160 non-null object
Vital registration complete                      111 non-null object
Latest agricultural census                       142 non-null object
Latest industrial data                           107 non-null float64
Latest trade data                                185 non-null float64
Latest water withdrawal data                     179 non-null object
Unnamed: 31                                      0 non-null float64
dtypes: float64(4), object(28)
memory usage: 60.4+ KB
```

EdStatsCountry-Series.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 613 entries, 0 to 612
Data columns (total 4 columns):
CountryCode     613 non-null object
SeriesCode      613 non-null object
DESCRIPTION     613 non-null object
Unnamed: 3      0 non-null float64
dtypes: float64(1), object(3)
memory usage: 19.3+ KB
```

REDONDANT

EdStatsData.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 70 columns):
Country Name        886930 non-null object
Country Code        886930 non-null object
Indicator Name      886930 non-null object
Indicator Code      886930 non-null object
1970                72288 non-null float64
1971                35537 non-null float64
1972                35619 non-null float64
1973                35545 non-null float64
1974                35730 non-null float64
1975                87306 non-null float64
1976                37483 non-null float64
1977                37574 non-null float64
```

EdStatsSeries.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Data columns (total 21 columns):
Series Code                     3665 non-null object
Topic                           3665 non-null object
Indicator Name                  3665 non-null object
Short definition                2156 non-null object
Long definition                 3665 non-null object
Unit of measure                 0 non-null float64
Periodicity                     99 non-null object
Base Period                     314 non-null object
Other notes                     552 non-null object
Aggregation method              47 non-null object
Limitations and exceptions      14 non-null object
Notes from original source      0 non-null float64
General comments                14 non-null object
Source                          3665 non-null object
Statistical concept and methodology 23 non-null object
Development relevance           3 non-null object
Related source links            215 non-null object
Other web links                 0 non-null float64
Related indicators              0 non-null float64
License Type                    0 non-null float64
Unnamed: 20                     0 non-null float64
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```

EdStatsFootNote.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 643638 entries, 0 to 643637
Data columns (total 5 columns):
CountryCode     643638 non-null object
SeriesCode      643638 non-null object
Year            643638 non-null object
DESCRIPTION     643638 non-null object
Unnamed: 4      0 non-null float64
dtypes: float64(1), object(4)
memory usage: 24.6+ MB
```

REDONDANT

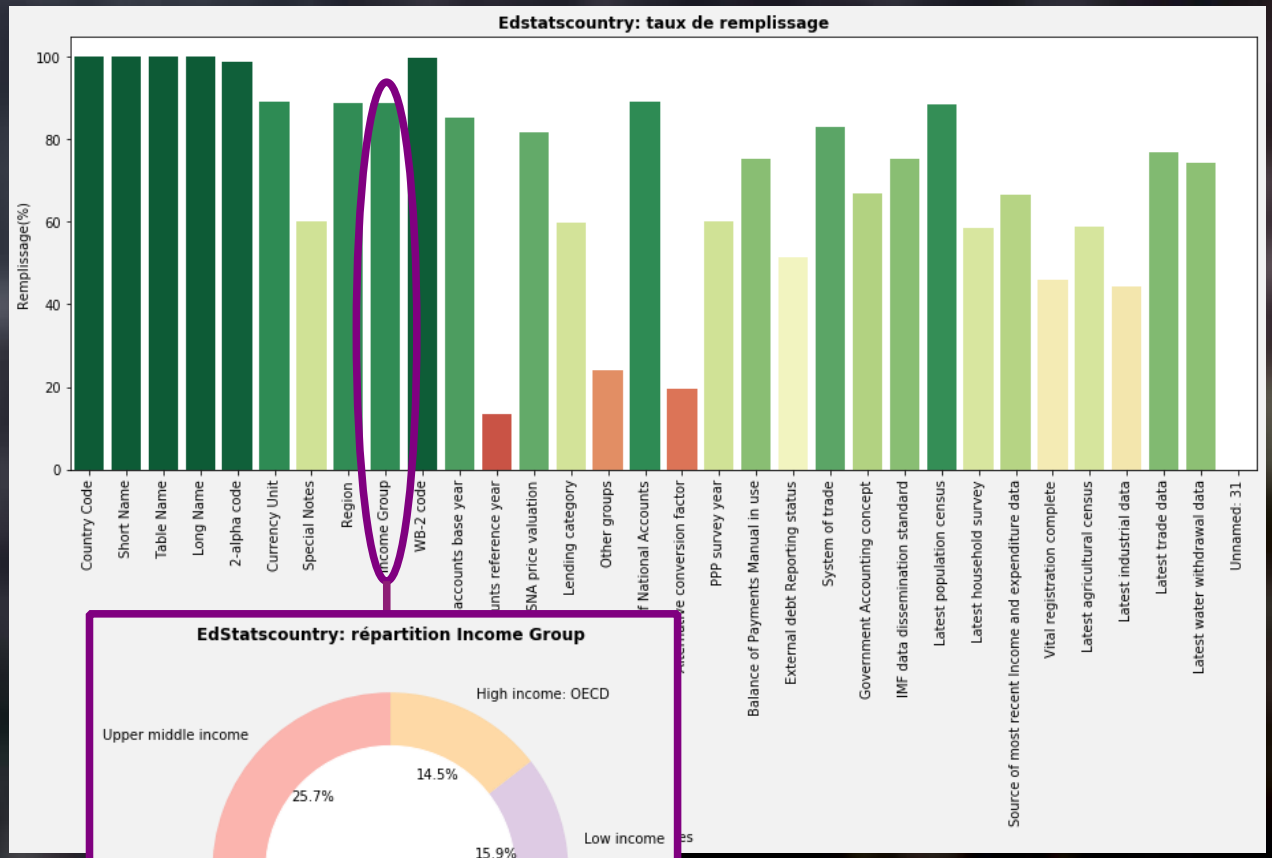
Contenu des datasets

Intéressant
Intéressant mais redondant/manquant

seaborn.barplot() :

EdStatsCountry.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241 entries, 0 to 240
Data columns (total 32 columns):
Country Code                                   241 non-null object
Short Name                                     241 non-null object
Table Name                                     241 non-null object
Long Name                                      241 non-null object
2-alpha code                                   238 non-null object
Currency Unit                                  215 non-null object
Special Notes                                  145 non-null object
Region                                         214 non-null object
Income Group                                   214 non-null object
WB-2 code                                      240 non-null object
National accounts base year                    205 non-null object
National accounts reference year               32 non-null float64
SNA price valuation                            197 non-null object
Lending category                               144 non-null object
Other groups                                   58 non-null object
System of National Accounts                    215 non-null object
Alternative conversion factor                  47 non-null object
PPP survey year                                145 non-null object
Balance of Payments Manual in use              181 non-null object
External debt Reporting status                 124 non-null object
System of trade                                200 non-null object
Government Accounting concept                  161 non-null object
IMF data dissemination standard                181 non-null object
Latest population census                       213 non-null object
Latest household survey                        141 non-null object
Source of most recent Income and expenditure data  160 non-null object
Vital registration complete                    111 non-null object
Latest agricultural census                     142 non-null object
Latest industrial data                         107 non-null float64
Latest trade data                              185 non-null float64
Latest water withdrawal data                   179 non-null object
Unnamed: 31                                    0 non-null float64
dtypes: float64(4), object(28)
memory usage: 60.4+ KB
```



Contenu des datasets

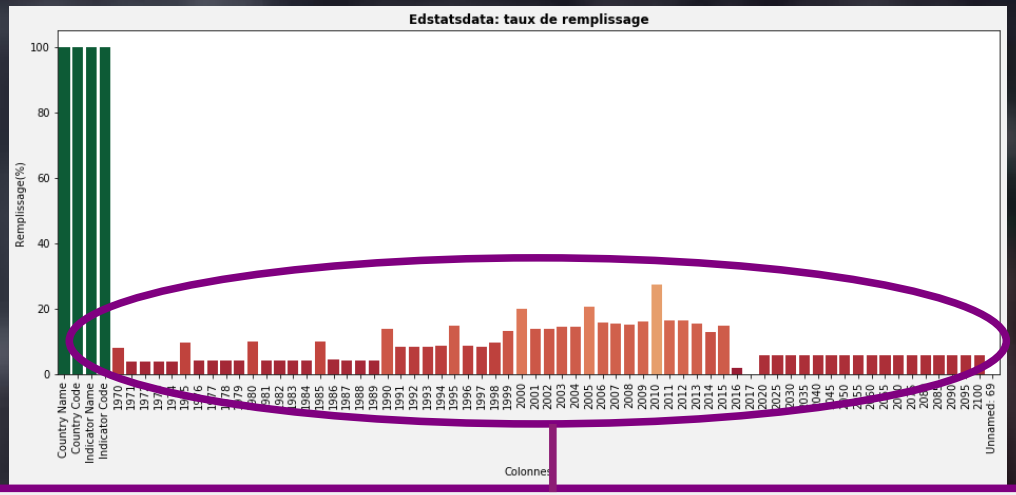
seaborn.barplot() :

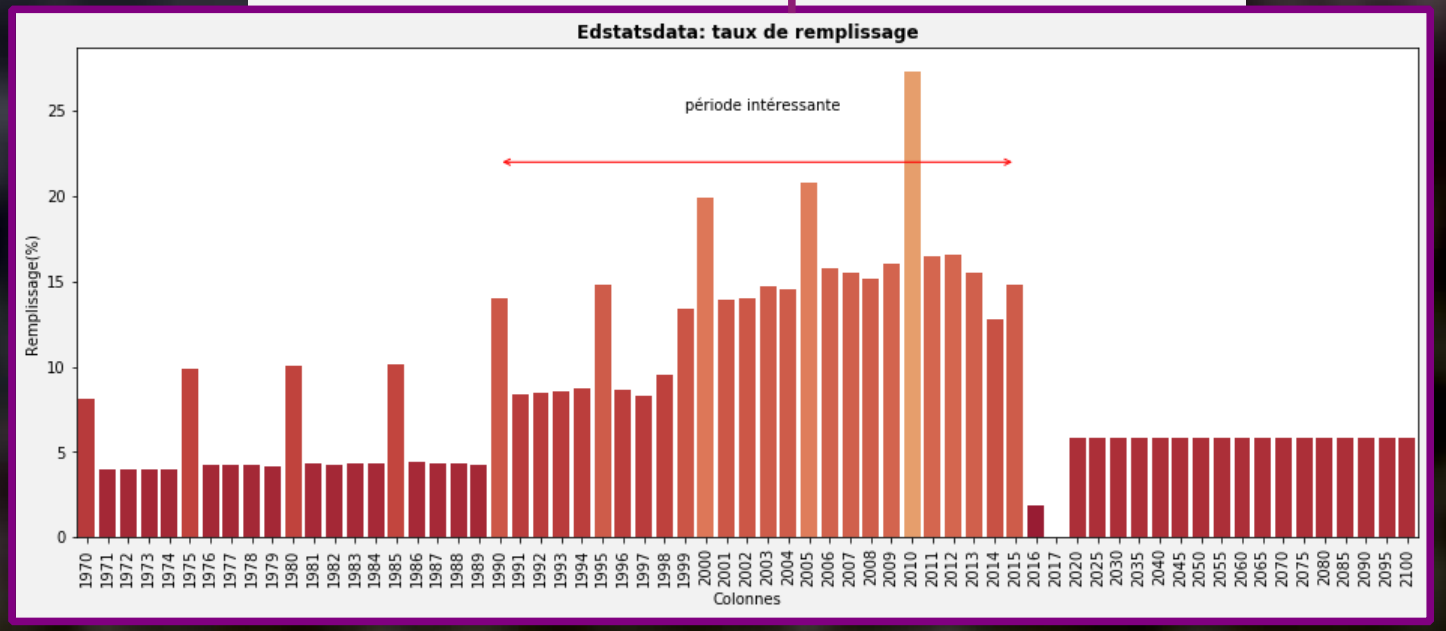
EdStatsData.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 70 columns):
Country Name        886930 non-null object
Country Code        886930 non-null object
Indicator Name      886930 non-null object
Indicator Code      886930 non-null object
1970                72288 non-null float64
1971                35537 non-null float64
1972                35619 non-null float64
1973                35545 non-null float64
1974                35730 non-null float64
1975                87306 non-null float64
1976                37483 non-null float64
1977                37574 non-null float64
```

Intéressant

Intéressant mais redondant/manquant





Contenu des datasets

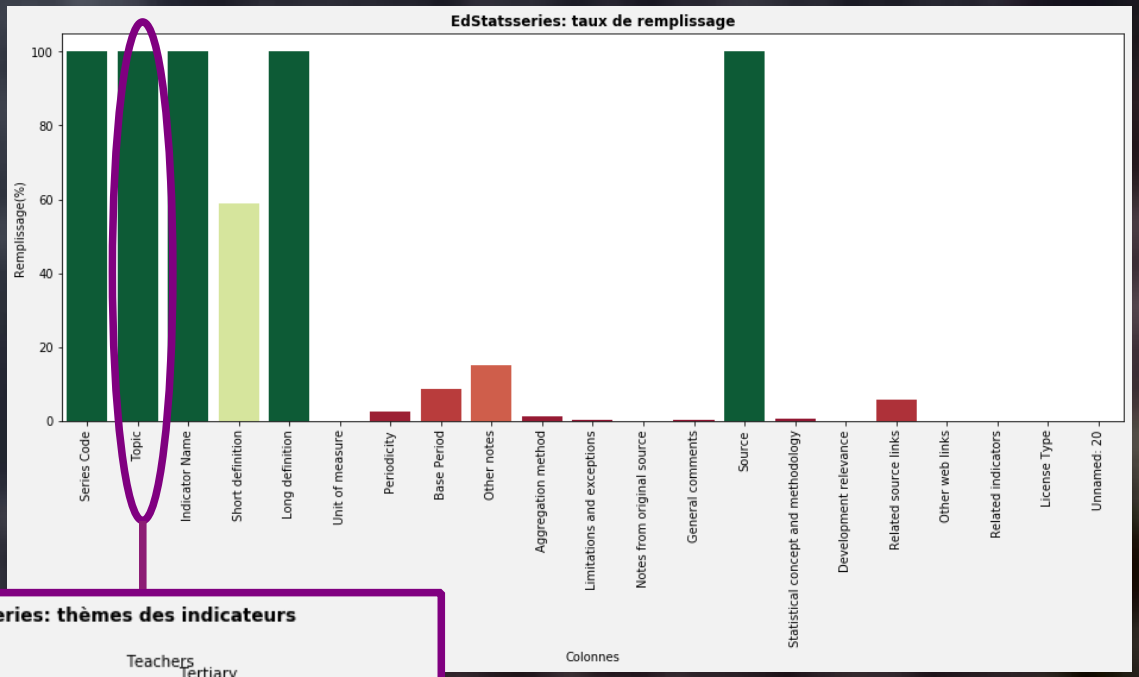
seaborn.barplot() :

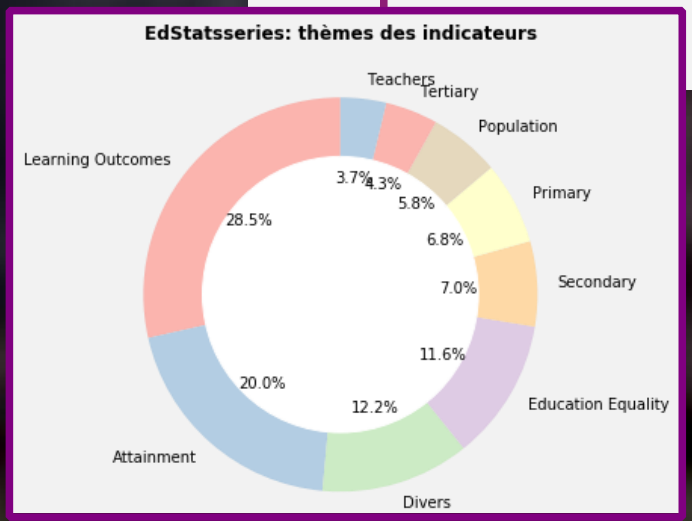
EdStatsSeries.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Data columns (total 21 columns):
Series Code                       3665 non-null object
Topic                             3665 non-null object
Indicator Name                    3665 non-null object
Short definition                  2156 non-null object
Long definition                   3665 non-null object
Unit of measure                   0 non-null float64
Periodicity                       99 non-null object
Base Period                       314 non-null object
Other notes                       552 non-null object
Aggregation method                47 non-null object
Limitations and exceptions        14 non-null object
Notes from original source        0 non-null float64
General comments                  14 non-null object
Source                            3665 non-null object
Statistical concept and methodology  23 non-null object
Development relevance             3 non-null object
Related source links              215 non-null object
Other web links                   0 non-null float64
Related indicators                0 non-null float64
License Type                      0 non-null float64
Unnamed: 20                       0 non-null float64
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```

Intéressant

Intéressant mais redondant/manquant





Nettoyage des données :

- Edition des datasets
- Fusion
- Elimination des doublons
- Subdivision en groupes

NETTOYAGE DES DONNÉES

Renommage de colonnes en *'CountryCode'* & *'IndicatorCode'* pour les fusions

Suppression de colonnes inutiles

Fusion des dataframes

Suppression des lignes nulles entre 1990 et 2015
+
Doublons
+
Indicateurs de projections

```
rename(columns= {'Country Code':'CountryC
drop(edstatsdata.loc[:, '1970':'1989']
merge(edstatsdata,on='IndicatorCode'
.dropna(subset=df.columns[8:34]
df[~df['IndicatorCode'].str.contains('PRJ')
df['Country Name'].uniqu
```

df

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 293570 entries, 610 to 870491
Data columns (total 37 columns):
CountryCode          293570 non-null object
Country Name         293570 non-null object
Region               284398 non-null object
Income Group         284398 non-null object
Topic                293570 non-null object
IndicatorCode        293570 non-null object
Indicator Name       293570 non-null object
Short definition     129533 non-null object
1990                 123000 non-null float64
1991                 72929 non-null float64
1992                 74189 non-null float64
```

dfa (tableau des régions)

dfc (tableau des pays)

dfi (tableau des income groups)

Suppression des indicateurs ne représentant pas la majorité des régions/pays/income groups

dfa : 5358 lignes, 38 cols

dfc : 219667 lignes, 38 cols

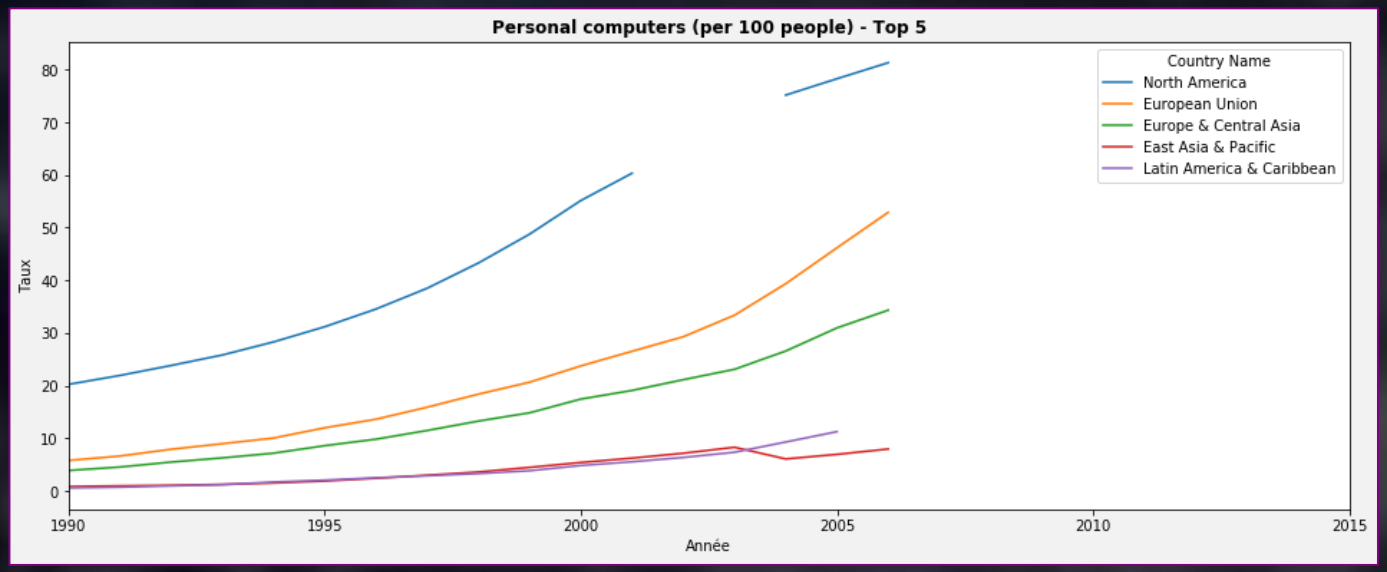
dfi : 3042 lignes, 38 cols

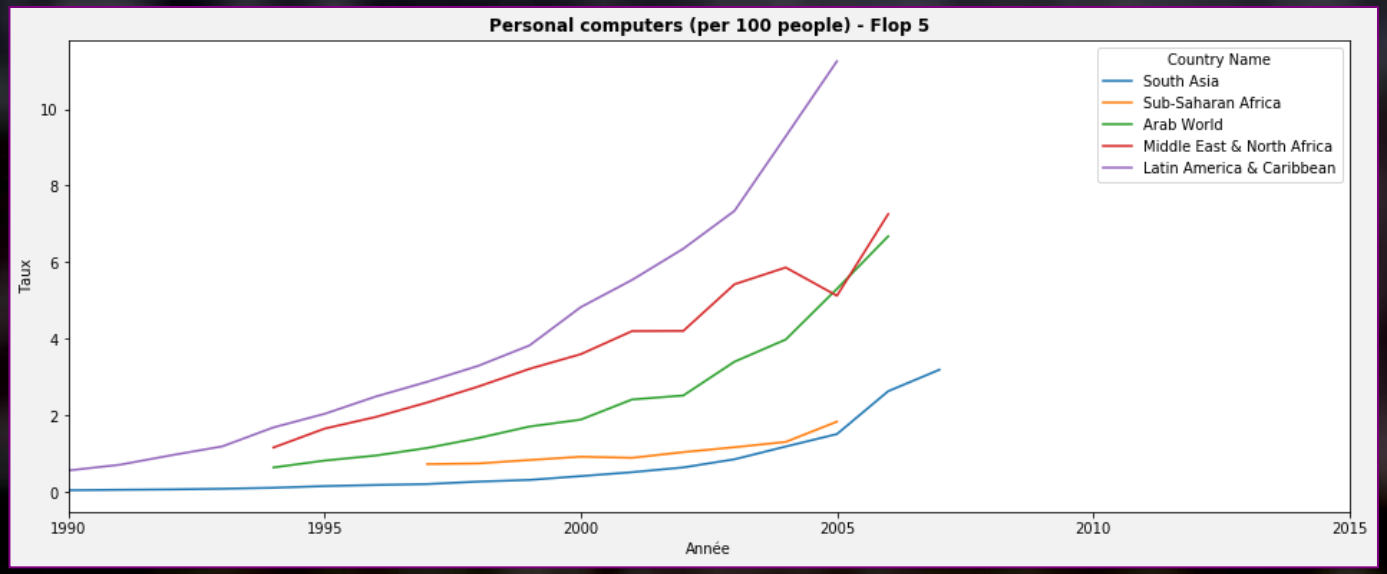
Infrastructures & communication :

- Ordinateurs personnels
- Utilisateurs d’Internet

ÉTUDE DE L'EXPANSION D'ACADEMY

Infrastructures & communication

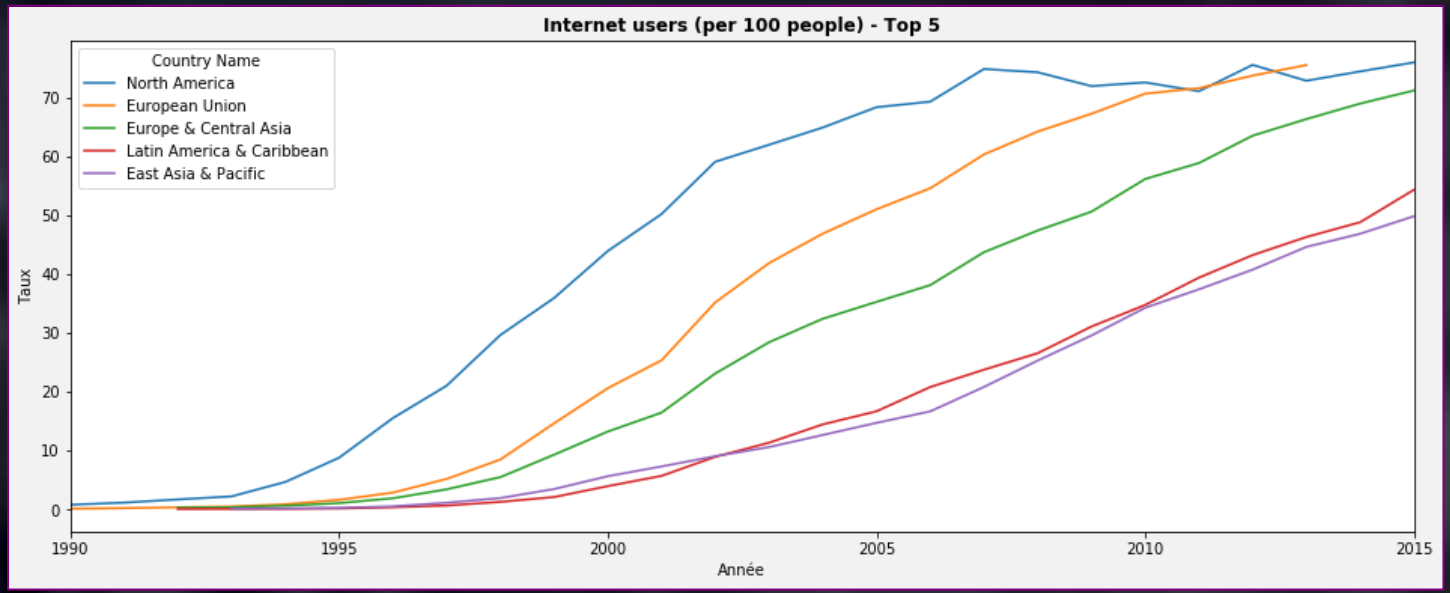


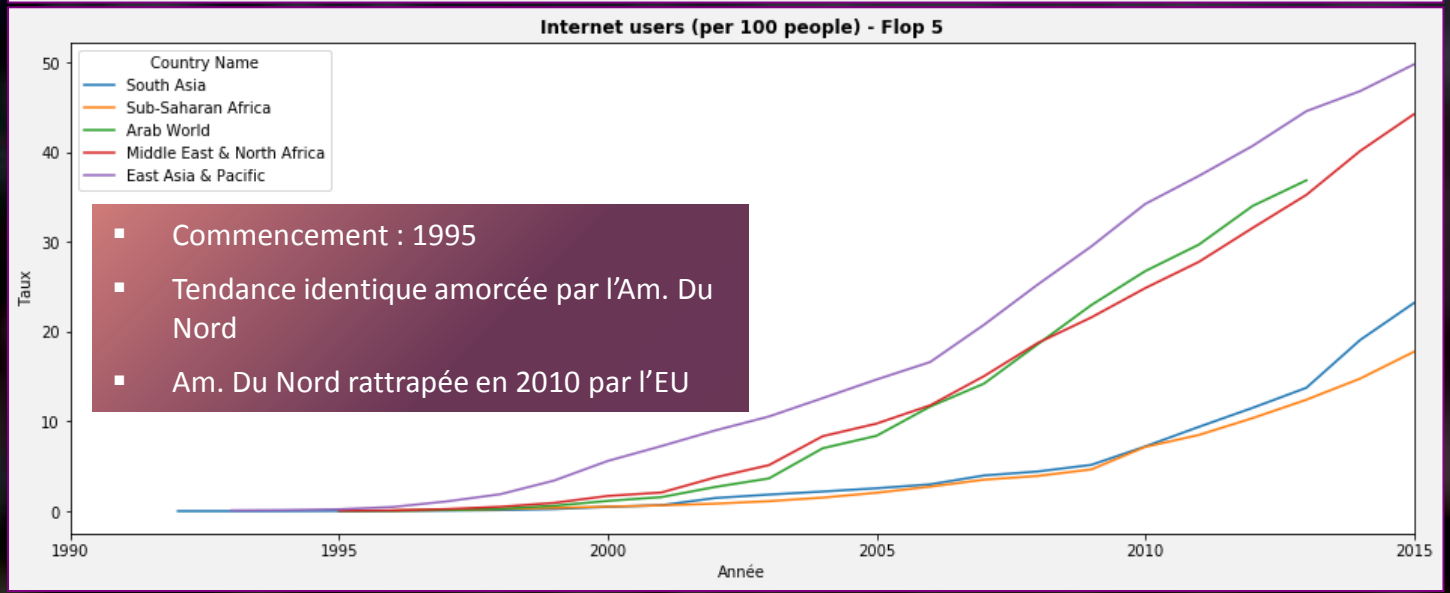


- 2005 + : interruption des données
- Même tendance
- Am. Nord reste la plus équipée
- Ecart important entre Am. Nord et reste du Monde
- Asie du Sud la moins équipée

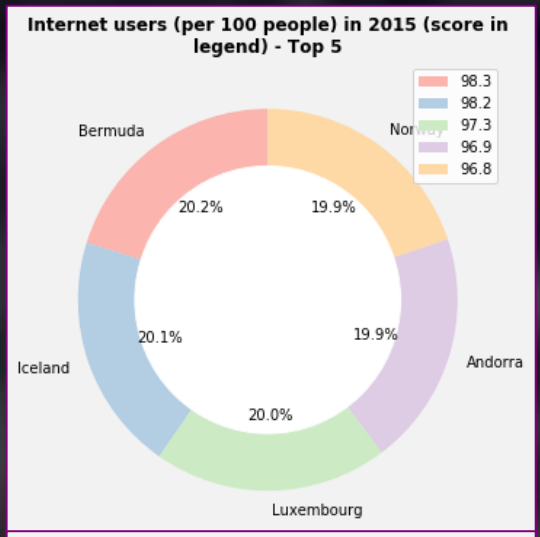
ÉTUDE DE L'EXPANSION D'ACADEMY

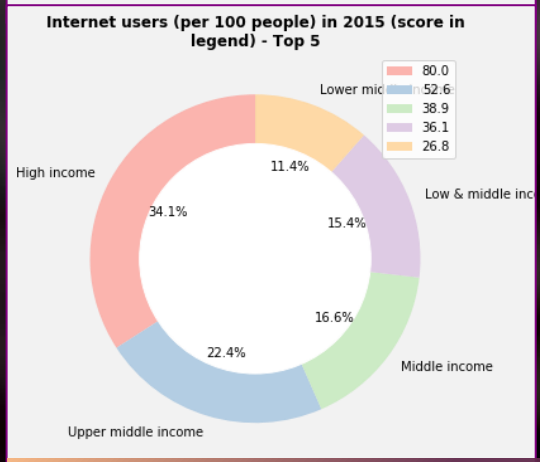
Infrastructures & communication





- Commencement : 1995
- Tendance identique amorcée par l'Am. Du Nord
- Am. Du Nord rattrapée en 2010 par l'EU





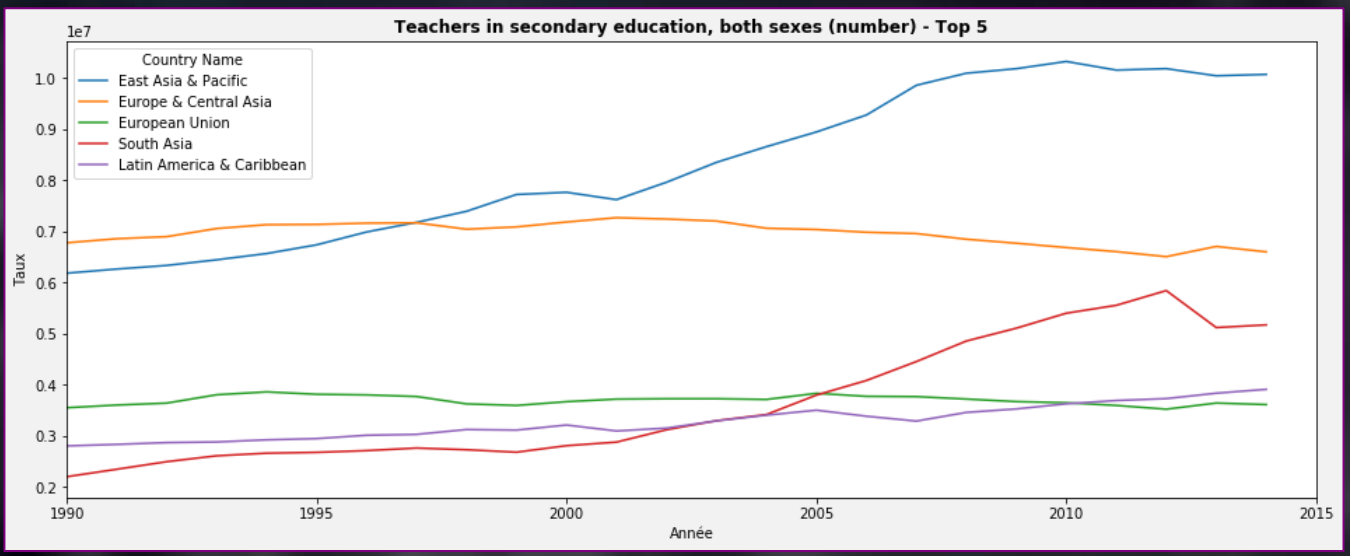
- Le Top 5 avoisine les 100%
- Les pays riches sont 3x plus équipés que les pays pauvres

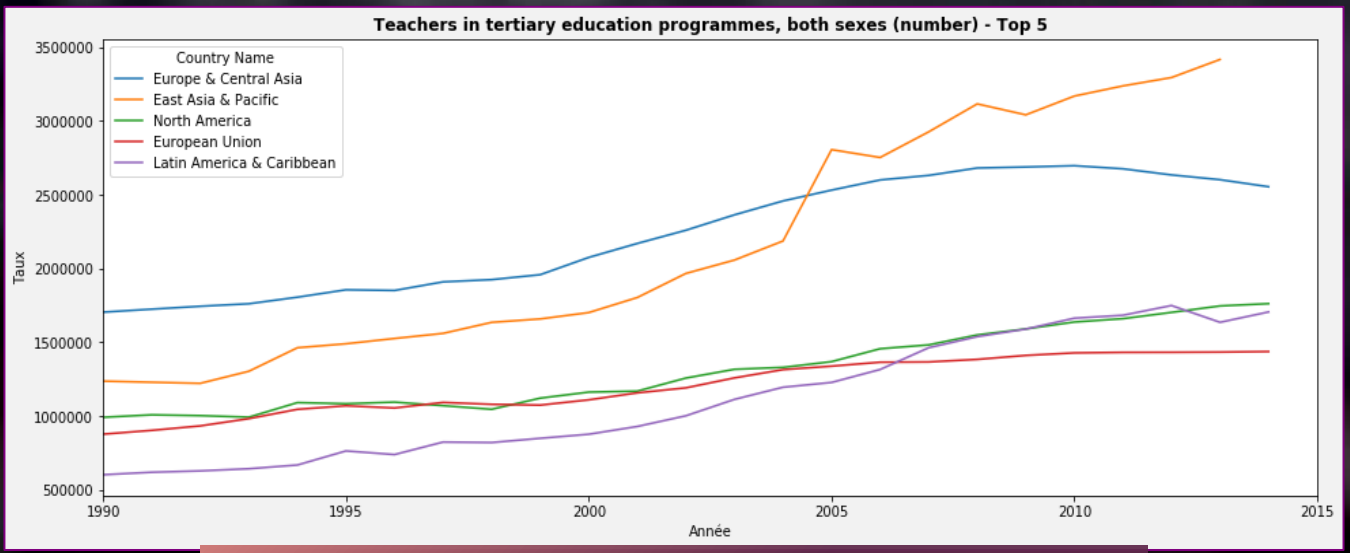
Enseignants :

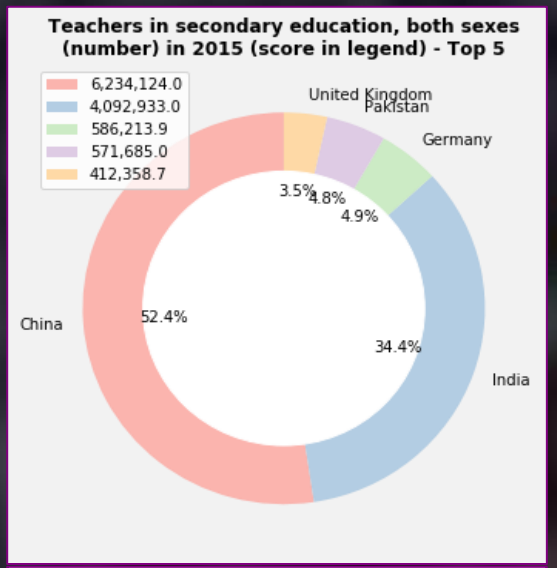
- Evolution des enseignants dans le secondaire
- Evolution des enseignants dans le tertiaire

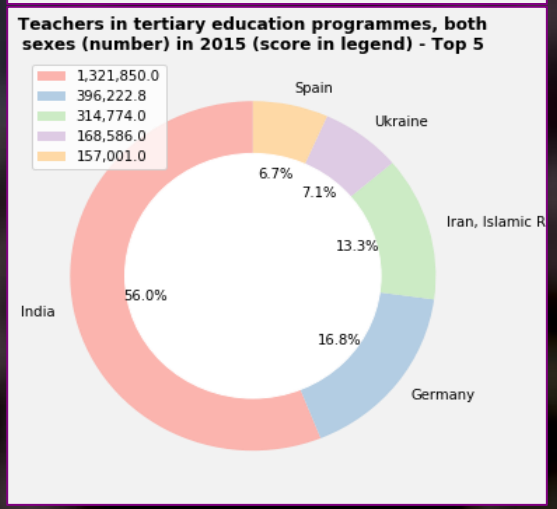
ÉTUDE DE L'EXPANSION D'ACADEMY

Enseignants









- 2003 : augmentation importante des enseignants en Asie dans le Secondaire & Tertiaire

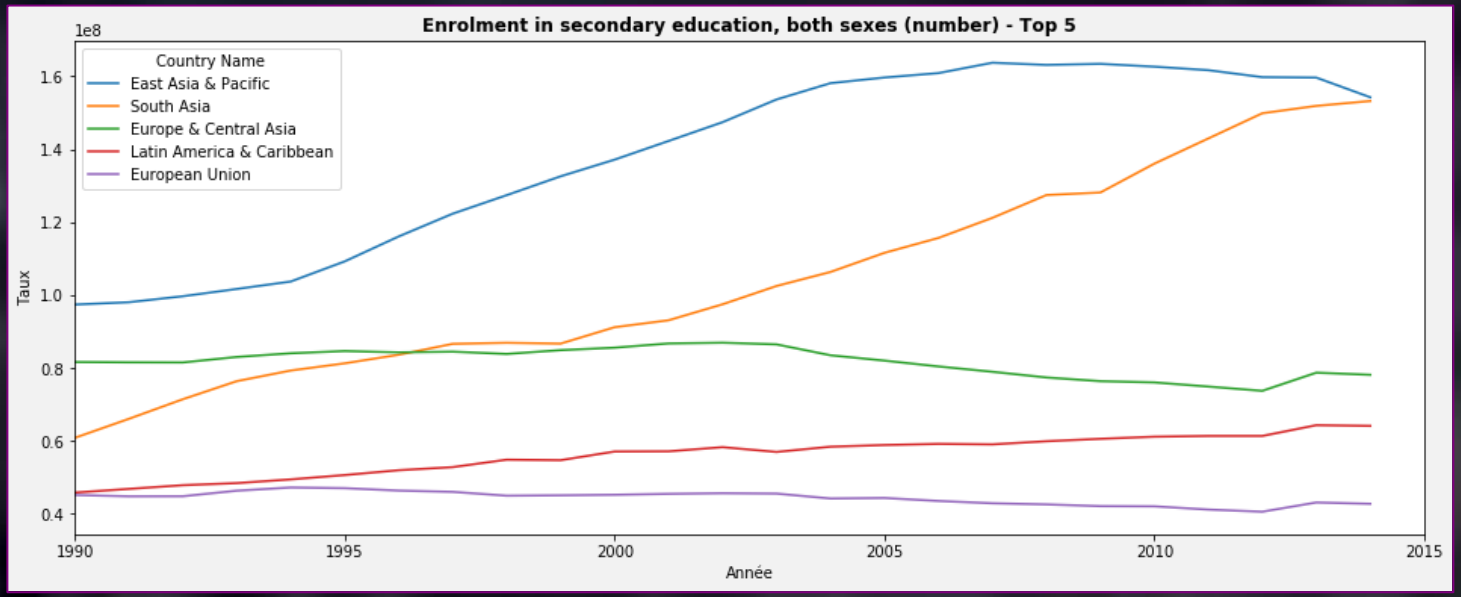
- La Chine domine le Secondaire
- L'Inde domine le Tertiaire

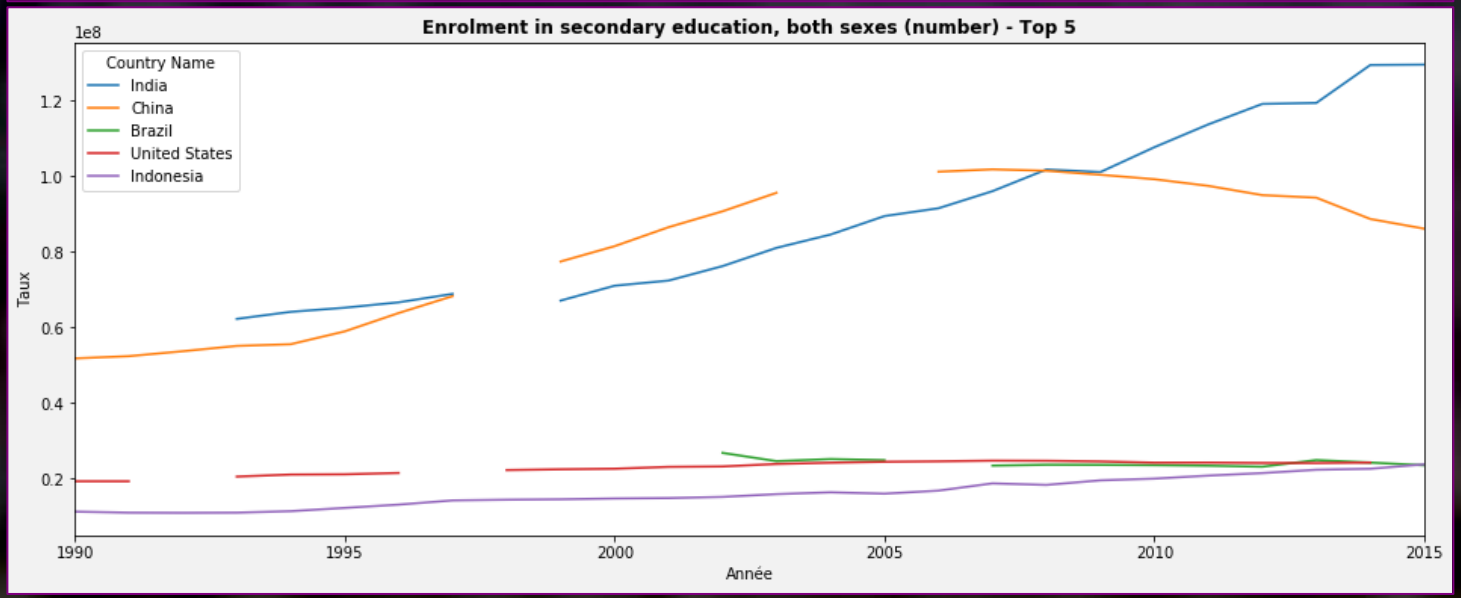
Etudes Secondaires :

- Inscriptions dans le secondaire
- Ecoles privées
- Redoublements
- Durées d'études

ÉTUDE DE L'EXPANSION D'ACADEMY

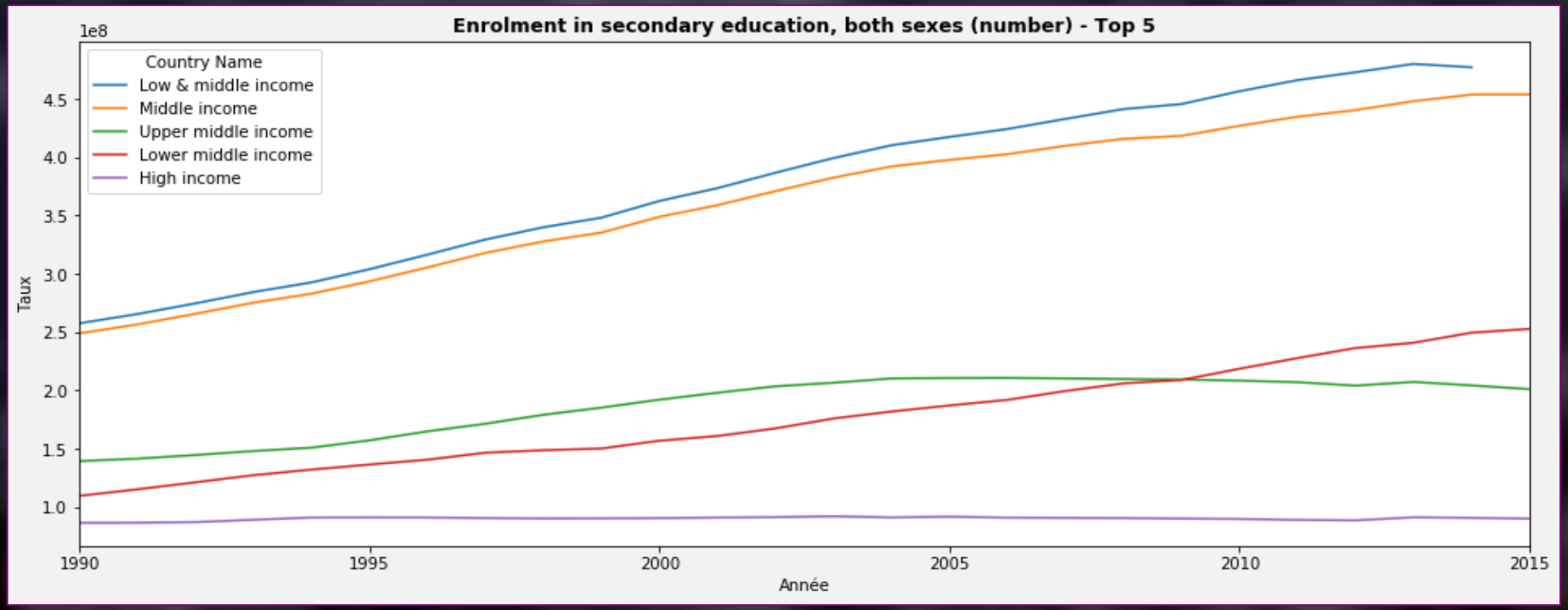
Etudes Secondaires





- Plusieurs pays ne renseignent qu'en périodes précises (recensement ?)
- 2015 : L'Asie atteint 160 M d'inscriptions
- Stagnation de l'Europe & Am. Latine
- Chine : apogée en 2007 puis baisse
- Inde : augmentation constante

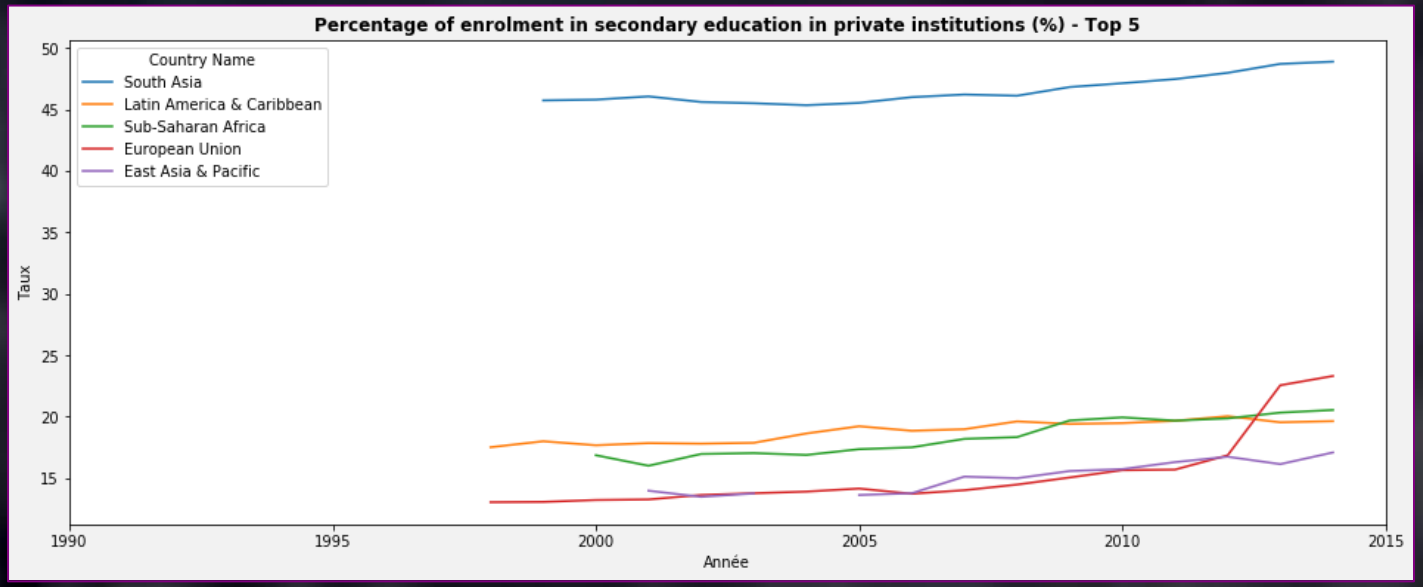
ÉTUDE DE L'EXPANSION D'ACADEMY

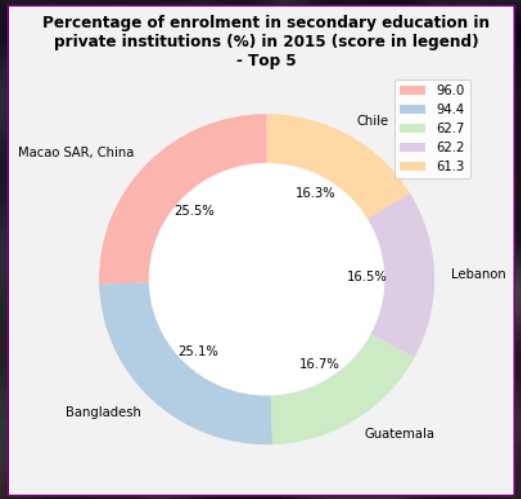


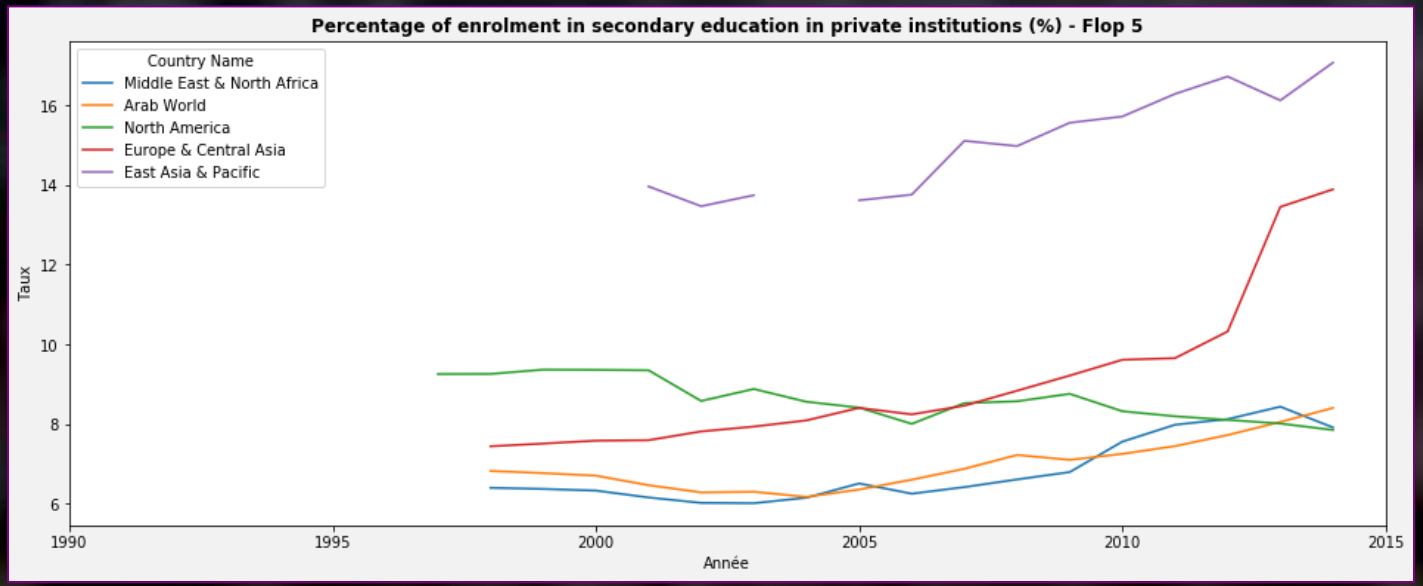
- Les pays à moyens revenus dominent
- Les pays à hauts revenus stagnent à 50 M d'inscrits

ÉTUDE DE L'EXPANSION D'ACADEMY

Etudes Secondaires



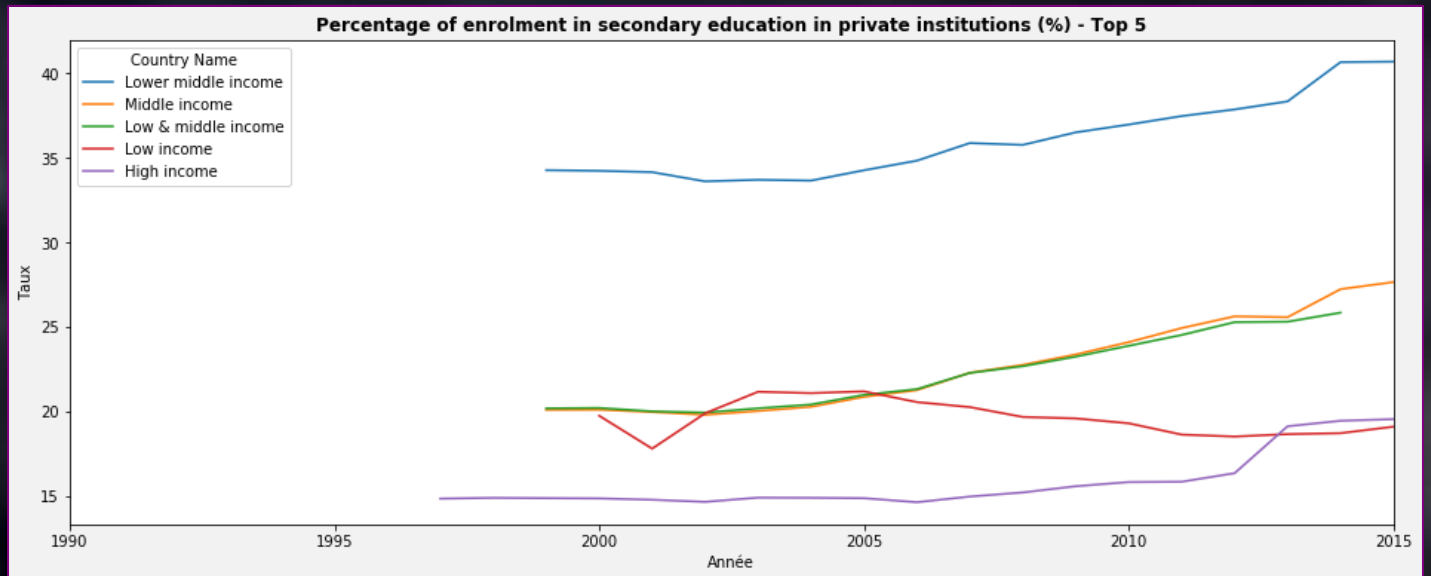


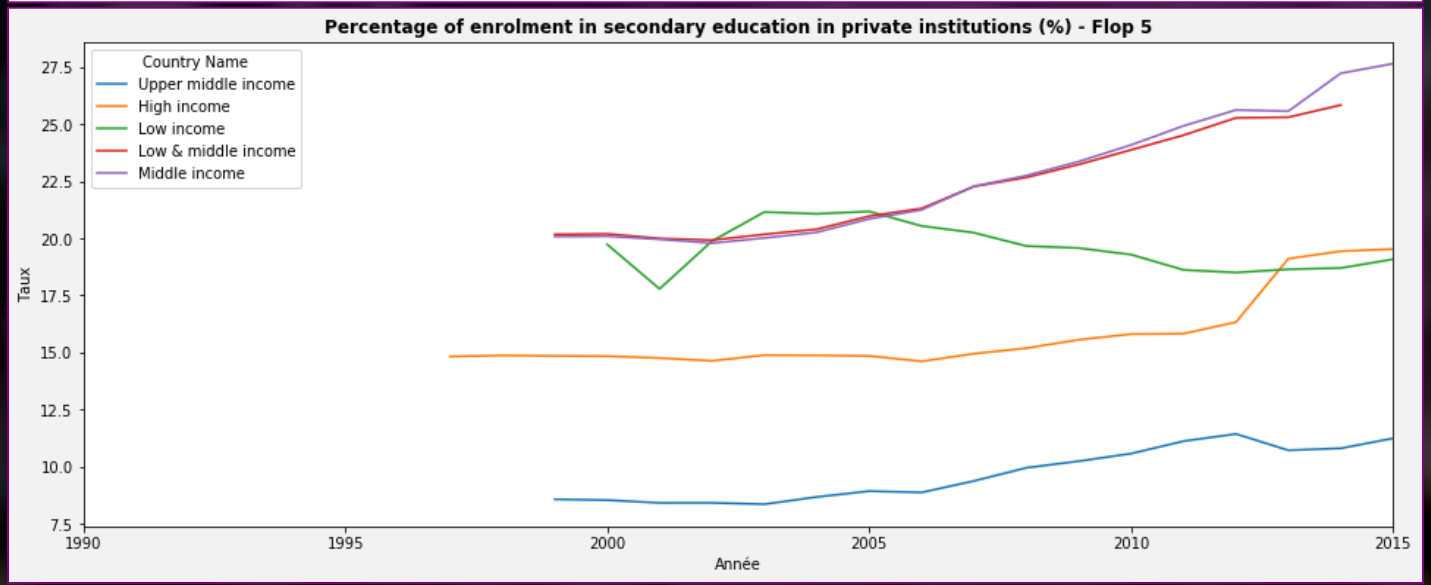


- Asie du Sud domine largement (~50% en 2015)
- Même tendance du Top 5
- Moyen Orient / Afrique du Nord / Monde Arabe : fréquentent le moins les établissements privés

ÉTUDE DE L'EXPANSION D'ACADEMY

Etudes Secondaires

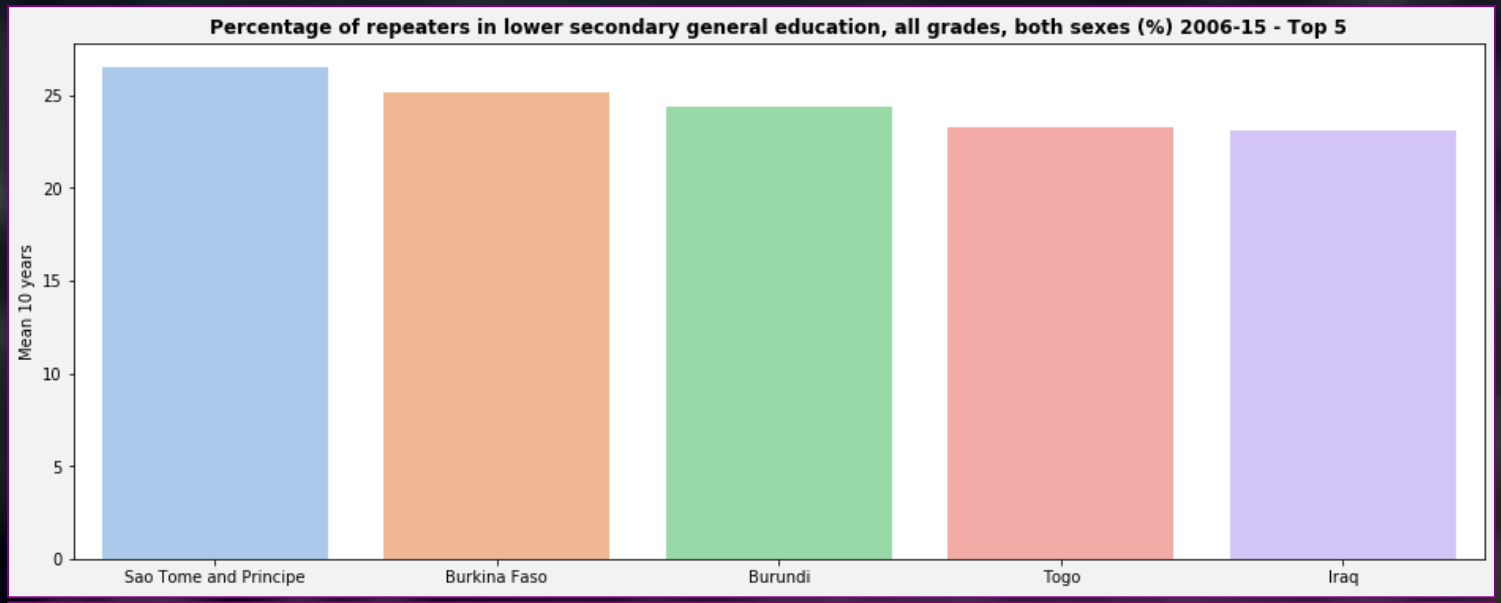


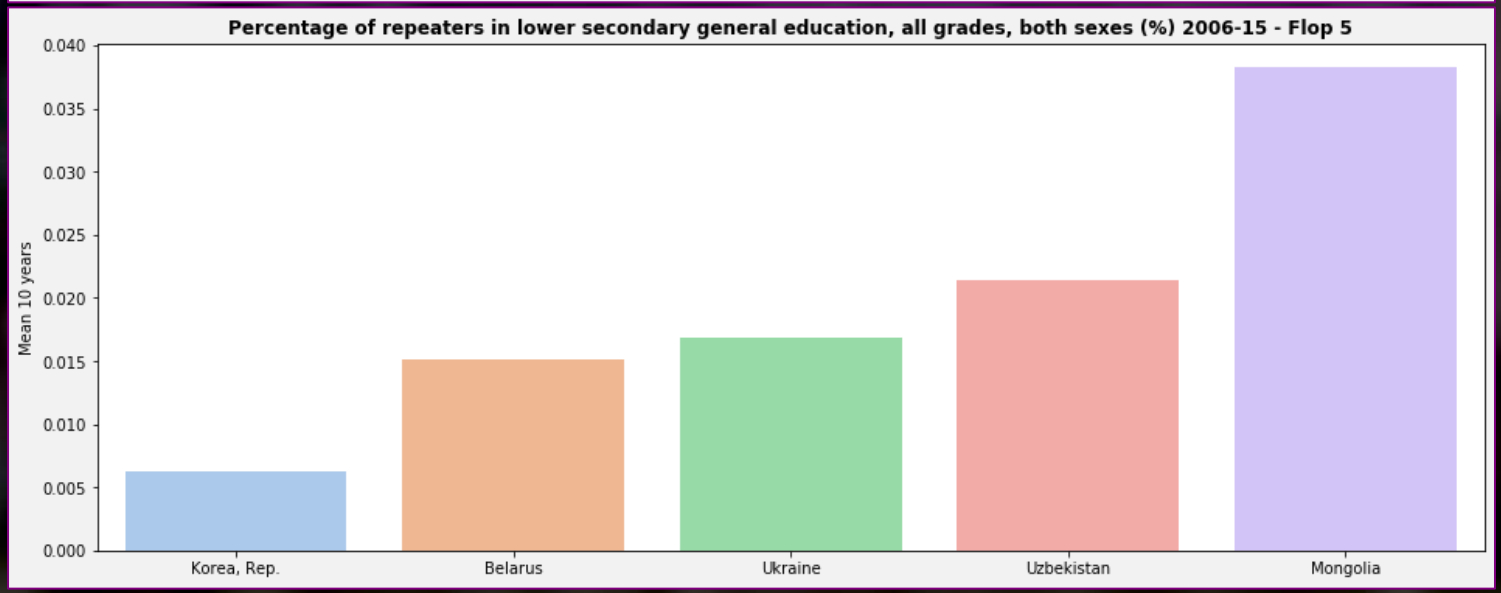


- L'ensemble des pays aux revenus moyens- / moyen / fréquentent le plus les établissements secondaires privés
- Les pays à revenus moyens- dominent largement les autres groupes (35% en 2000, 40% en 2015)
- Les pays riches ne dépassent pas les 20%

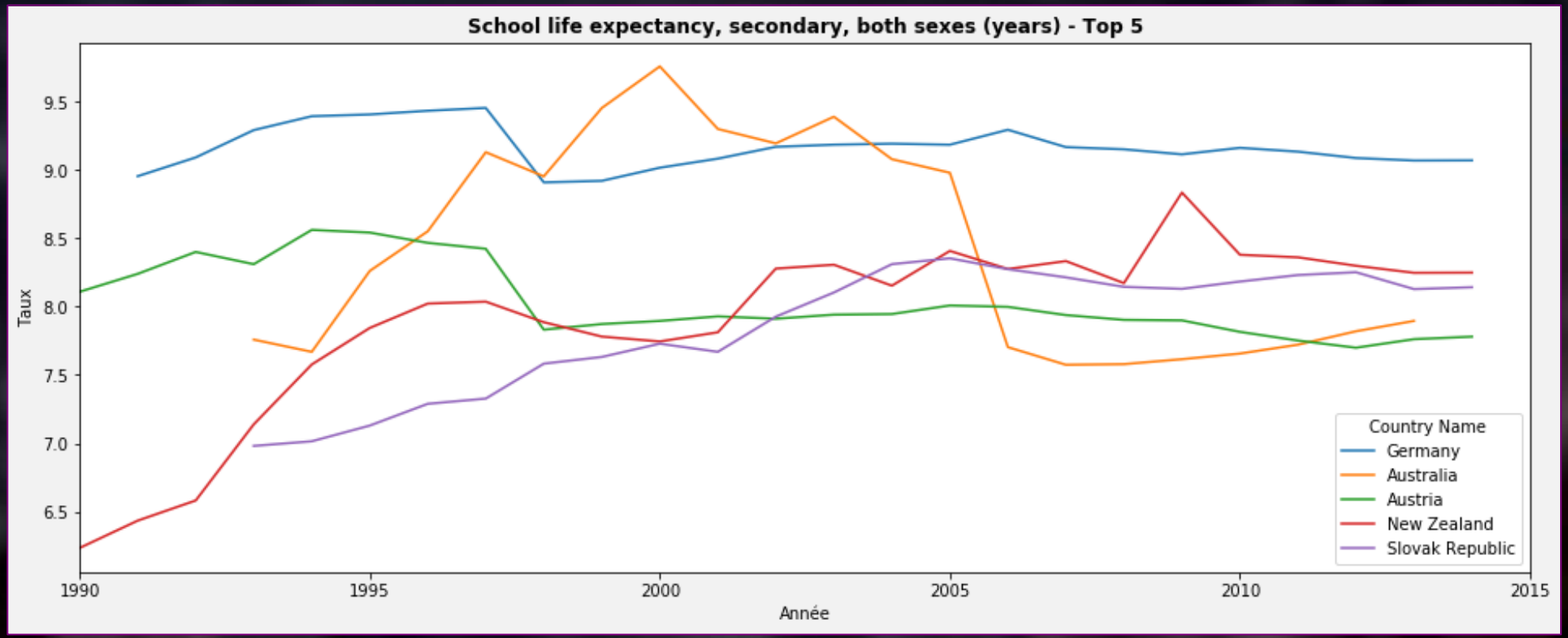
ÉTUDE DE L'EXPANSION D'ACADEMY

Etudes Secondaires



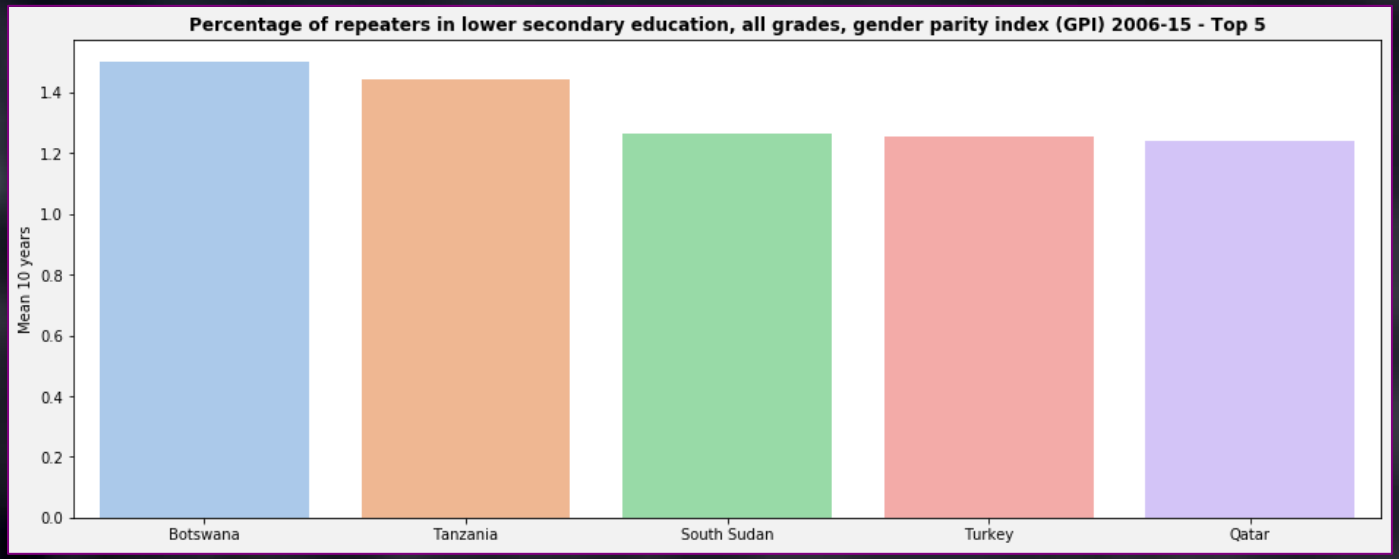


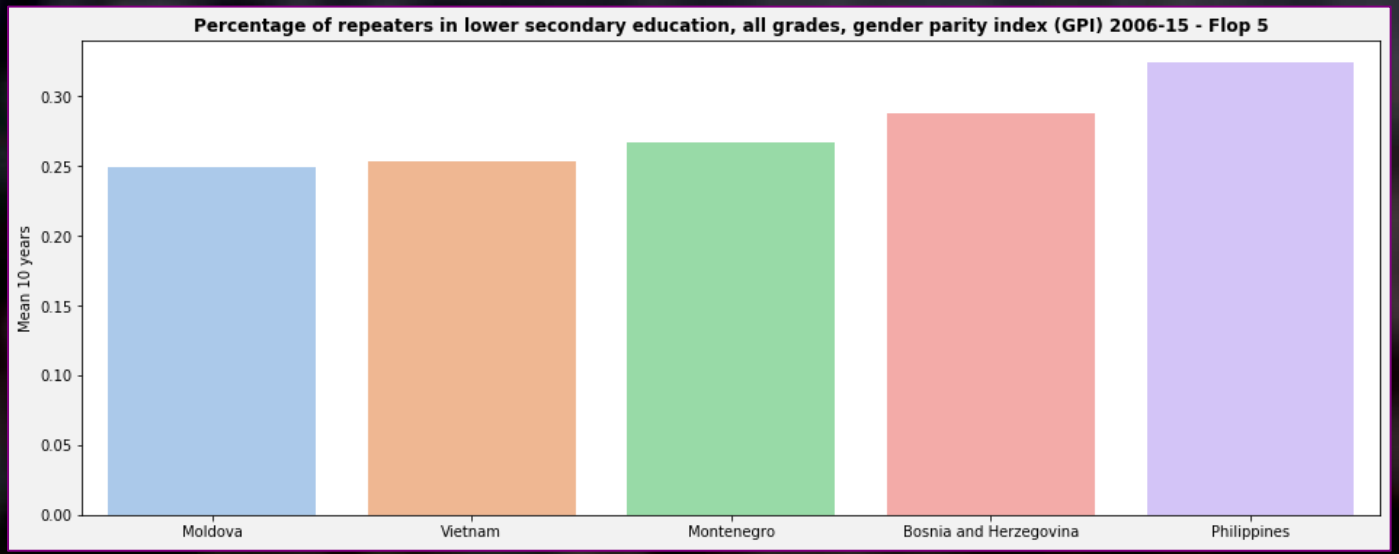
- Le plus faible = le meilleur
- ¼ de redoublants dans les pays du Top 5
- La Corée du Sud obtient le meilleur taux (quasi nul) de redoublants dans le secondaire



- La durée de scolarité dans le secondaire est de 9 ans en Allemagne (stable)
- Chute en Australie en 2005 (~ -1,5 an)

ÉTUDE DE L'EXPANSION D'ACADEMY





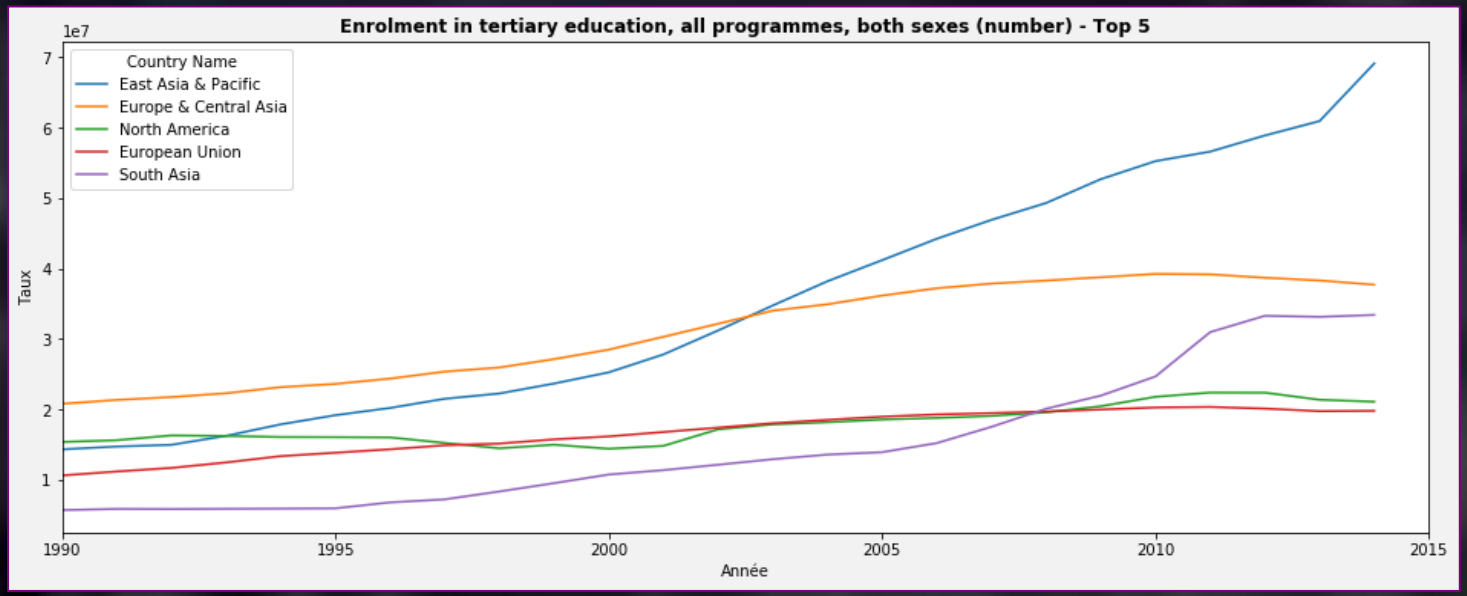
- GPI :
  - > 1 : domination féminine
  - = 1 : parité
  - < 1 : domination masculine
- Botswana & Tanzanie : les redoublants sont en majorité des femmes
- Moldavie & Vietnam : les redoublants sont en majorité des hommes

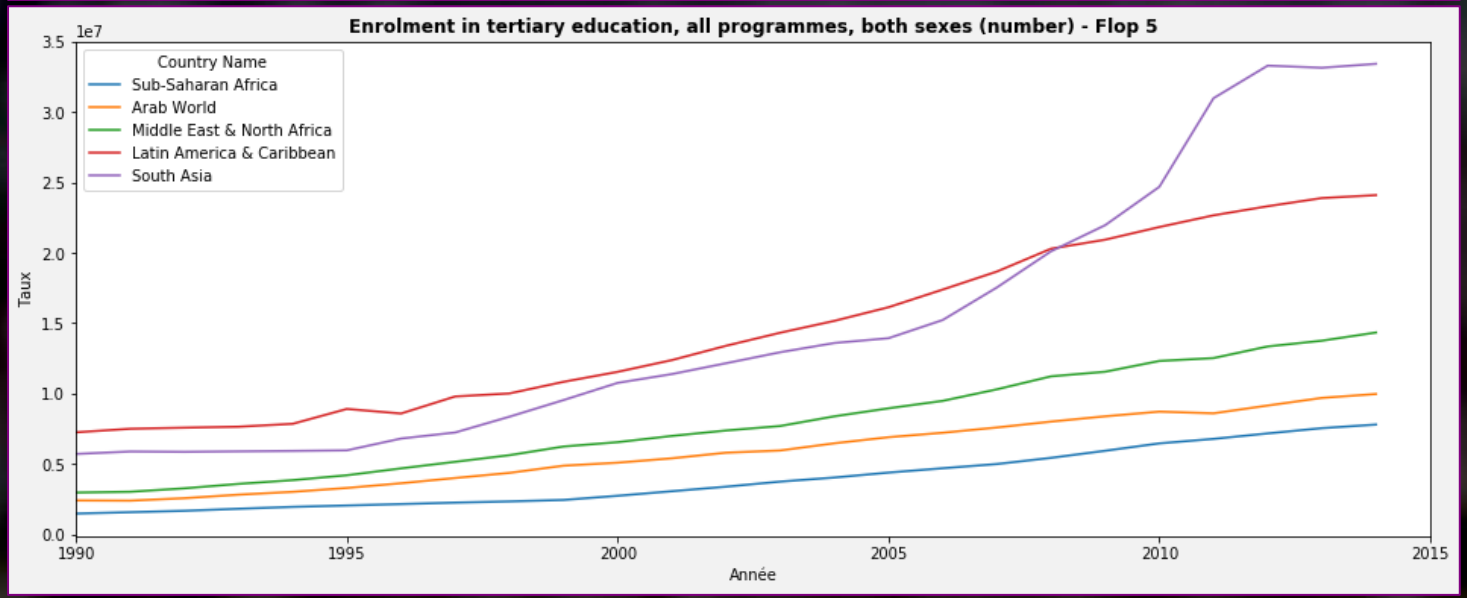
Etudes Tertiaires :

- Inscriptions dans le Tertiaire
- Ecoles privées
- Programmes
- Durées d'études

ÉTUDE DE L'EXPANSION D'ACADEMY

Etudes Tertiaires

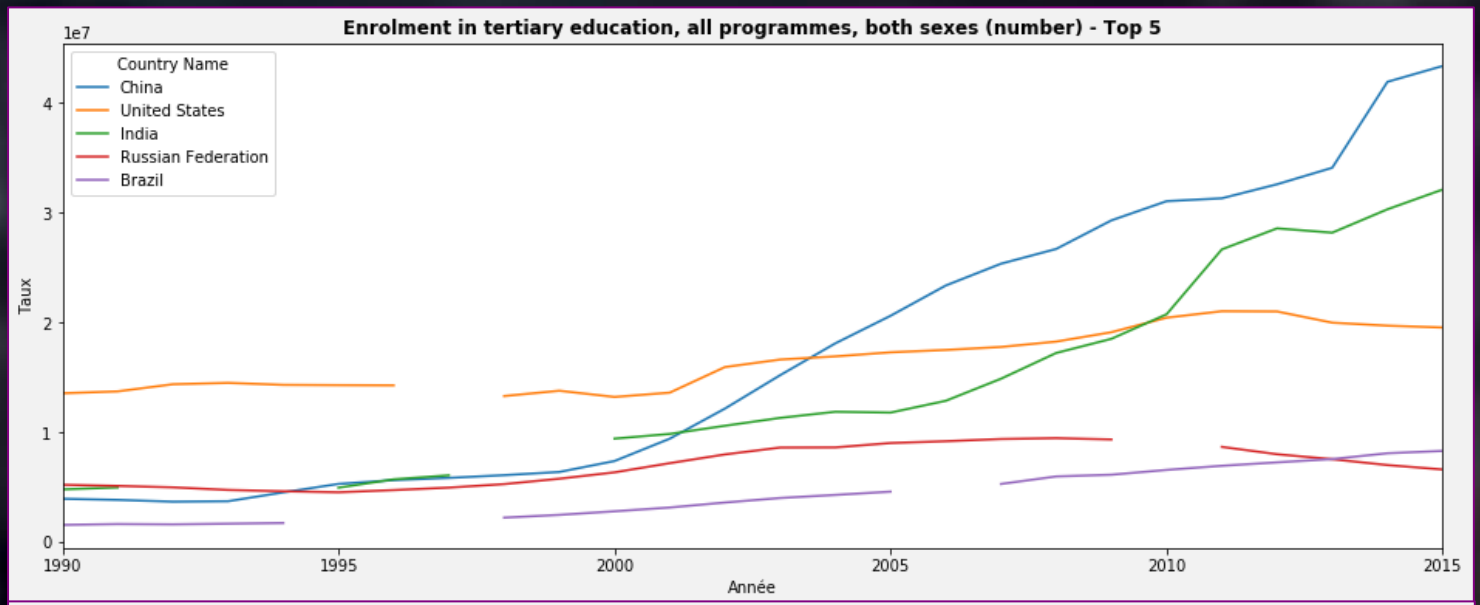


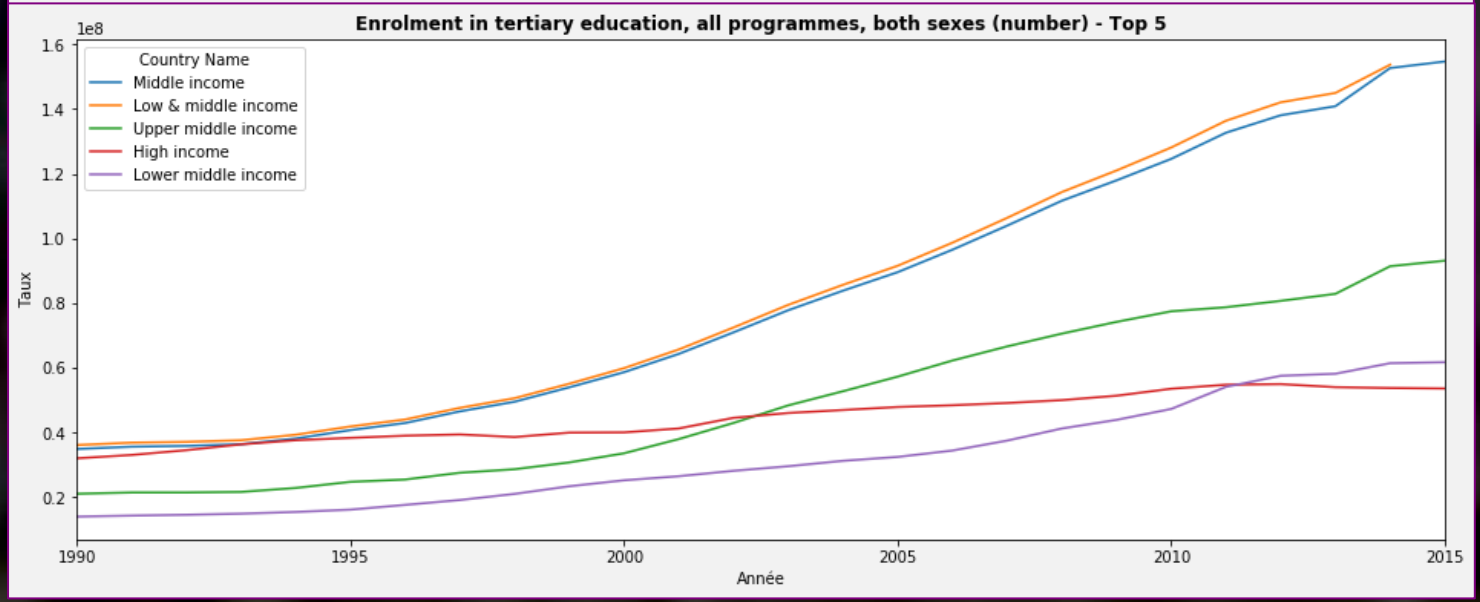


- 2015 : L'Asie de l'Est atteint 70 M d'inscriptions
- Afrique Sub-Saharienne suit la tendance tout en enregistrant le moins d'inscriptions

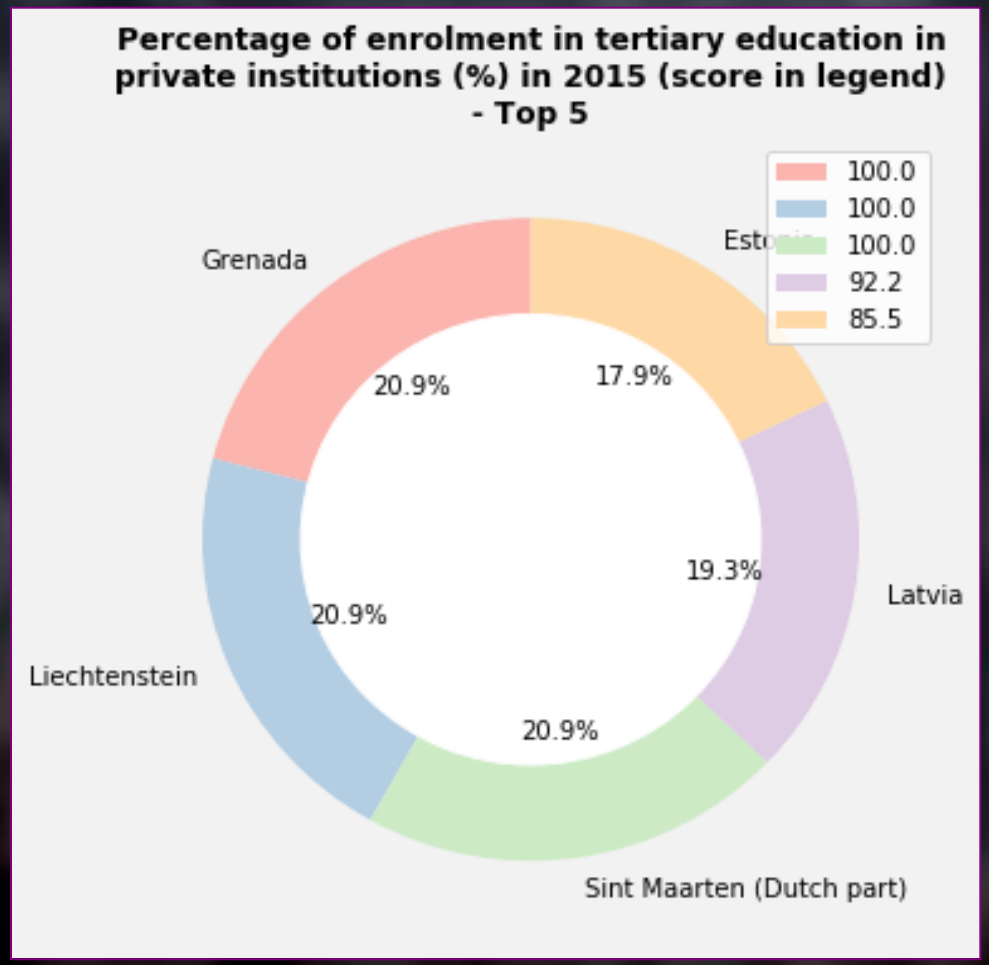
ÉTUDE DE L'EXPANSION D'ACADEMY

Etudes Tertiaires





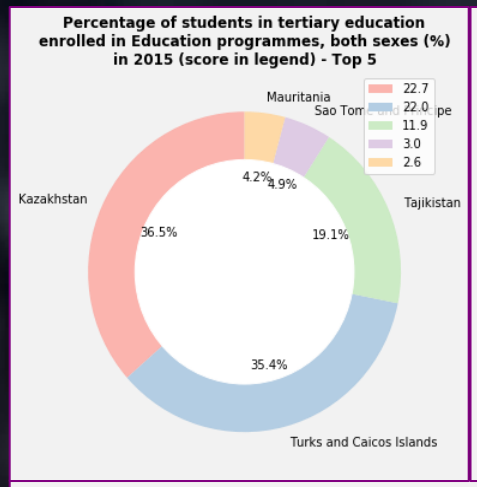
- Plusieurs pays ne renseignent qu'en périodes précises (recensement ?)
- 2003 : la Chine domine le reste du Monde
- L'Inde suit la même tendance, passe numéro 2 dès 2010
- Les pays à moyens revenus ont le plus d'étudiants inscrits dans le Tertiaire

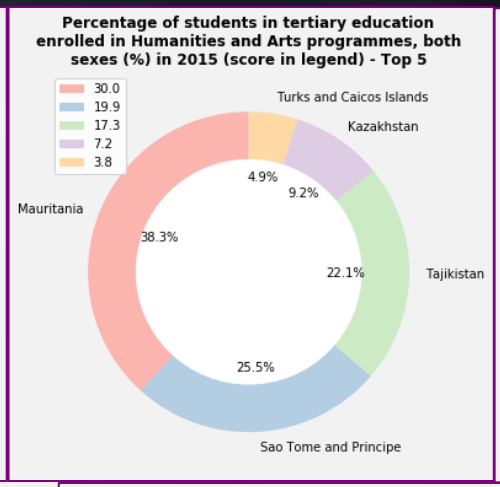


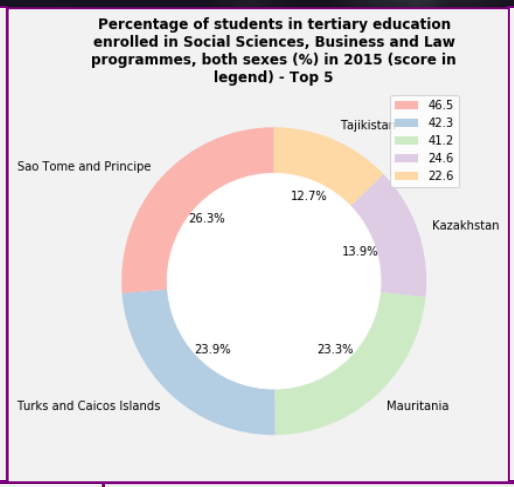
- 3 pays indiquent 100%
- Lettonie & Estonie avoisinent les 100%

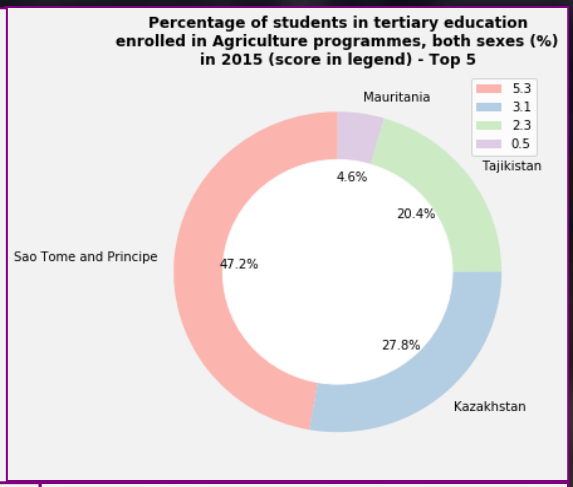
ÉTUDE DE L'EXPANSION D'ACADEMY

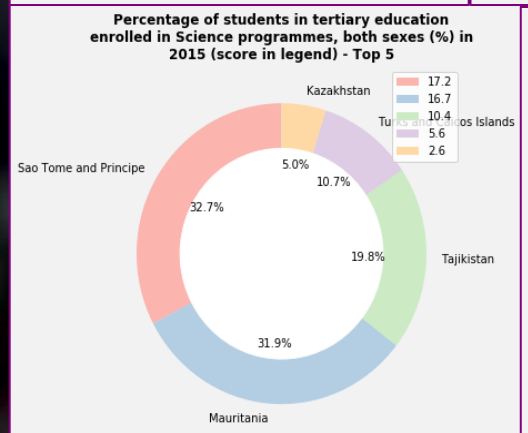
Etudes Tertiaires

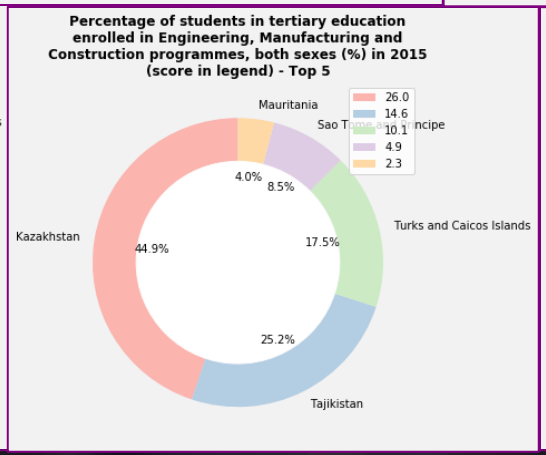


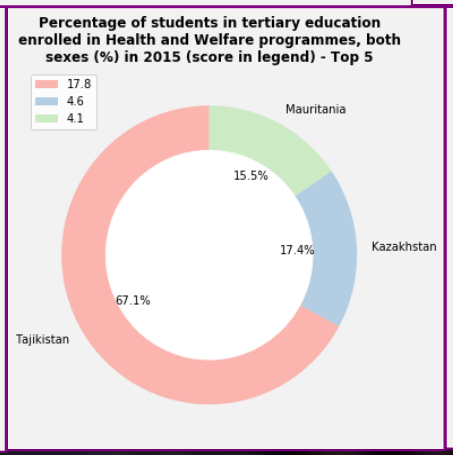


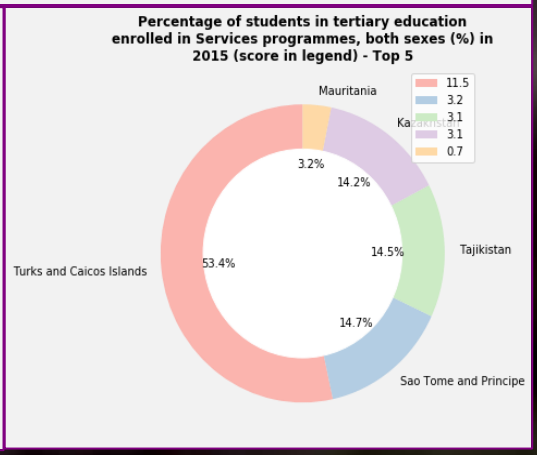








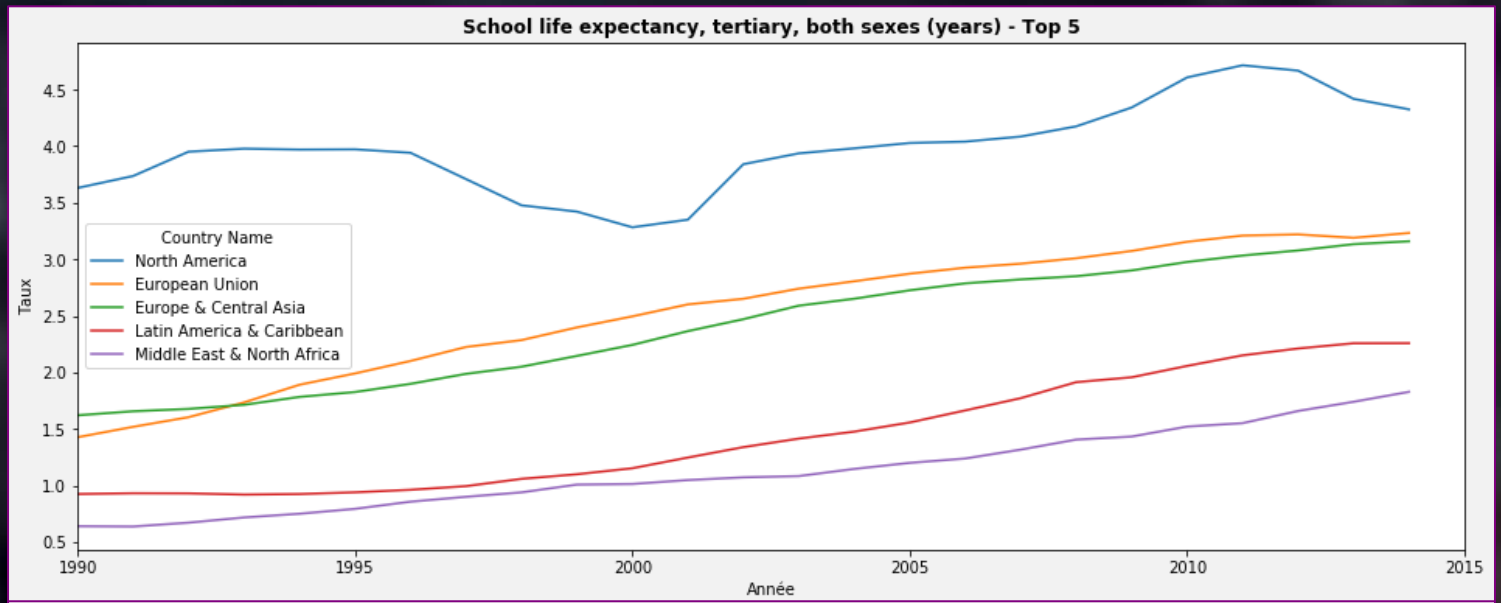


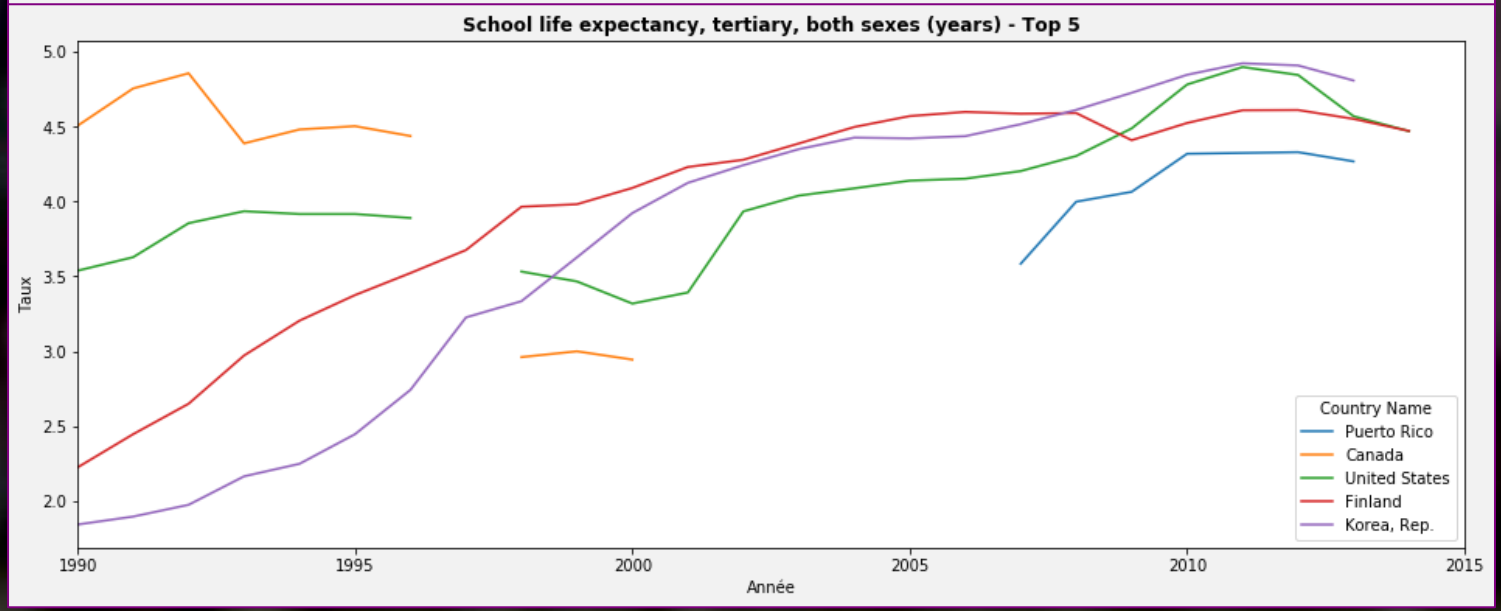


- Disparités dans les programmes d'Education / Sciences humaines & Art / Agriculture / Santé / Services
- Les Sciences Sociales / Economie / Droit atteignent jusqu'à 46% d'inscrits dans le Tertiaire

ÉTUDE DE L'EXPANSION D'ACADEMY

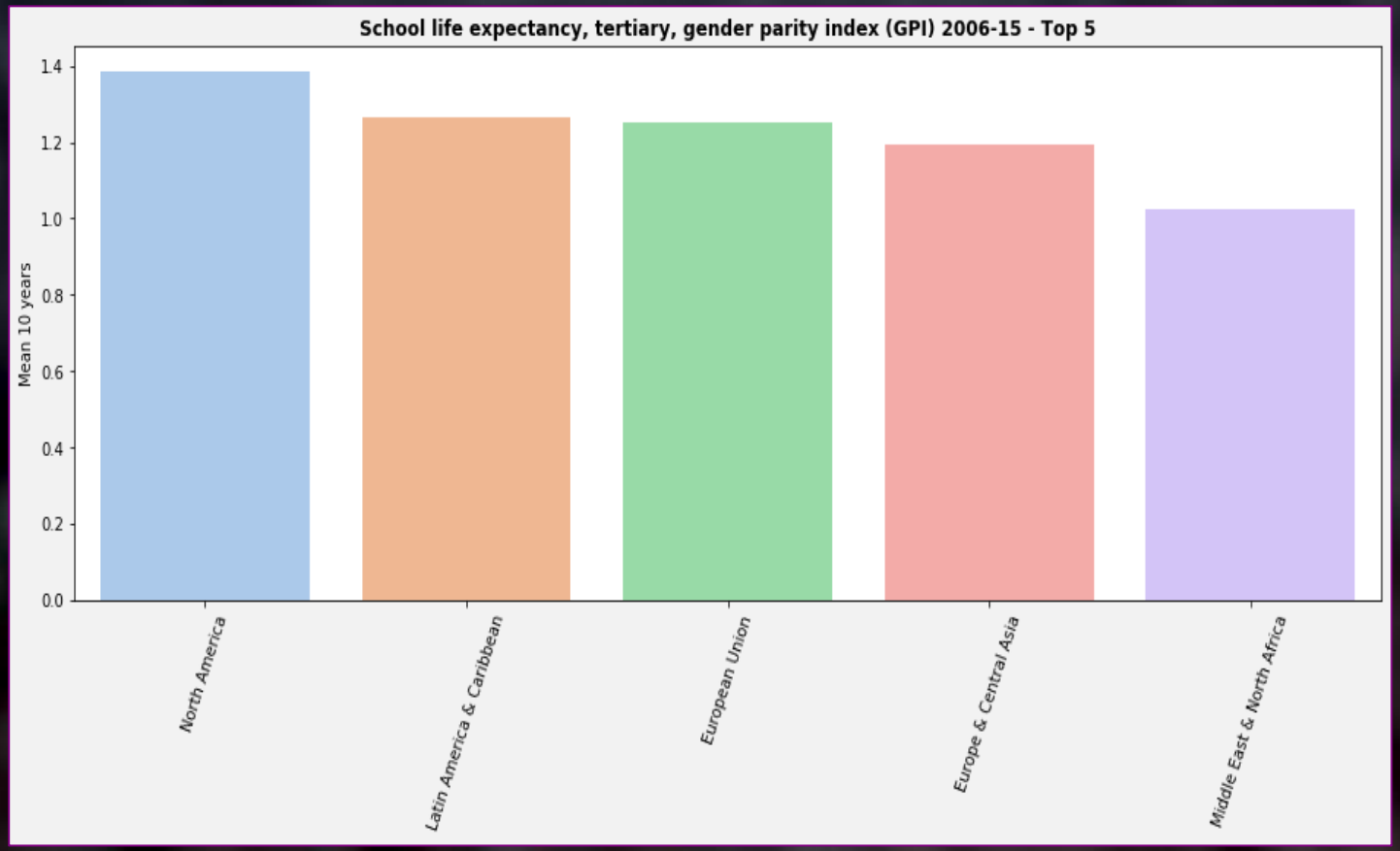
Etudes Tertiaires





- 2015 : 4,5 ans en Am. Du Nord, suivi par l'EU (3 ans)
- Corée du Sud & USA sont les plus longs
- Déclin des US à partir de 2011

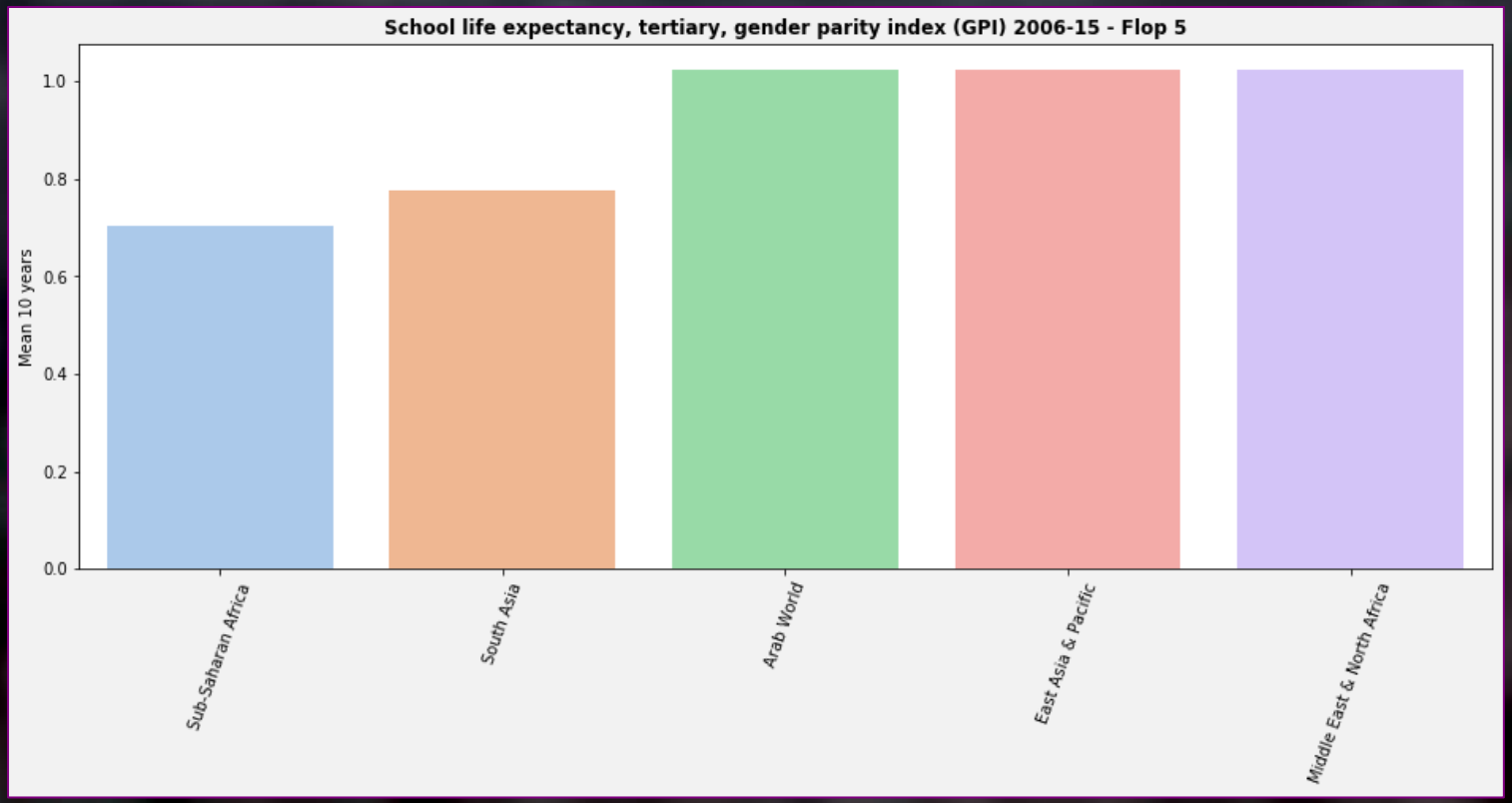
ÉTUDE DE L'EXPANSION D'ACADEMY



- GPI :
  - > 1 : domination féminine
  - = 1 : parité
  - < 1 : domination masculine
- Amérique : écoles supérieures terminées majoritairement par des femmes

ÉTUDE DE L'EXPANSION D'ACADEMY

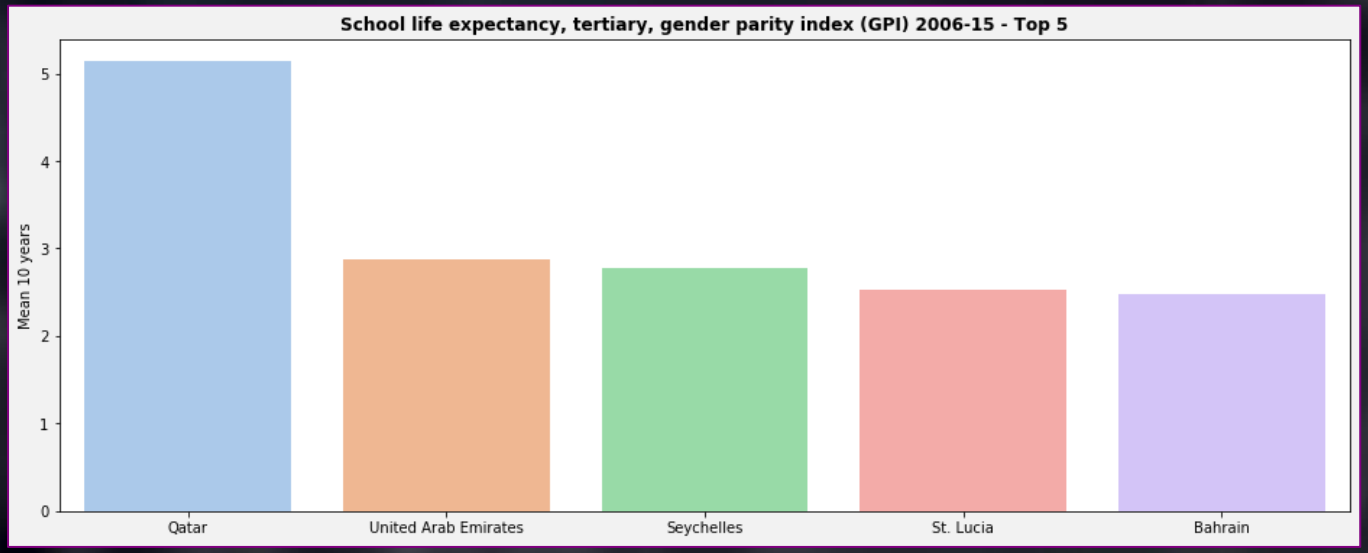
Etudes Tertiaires

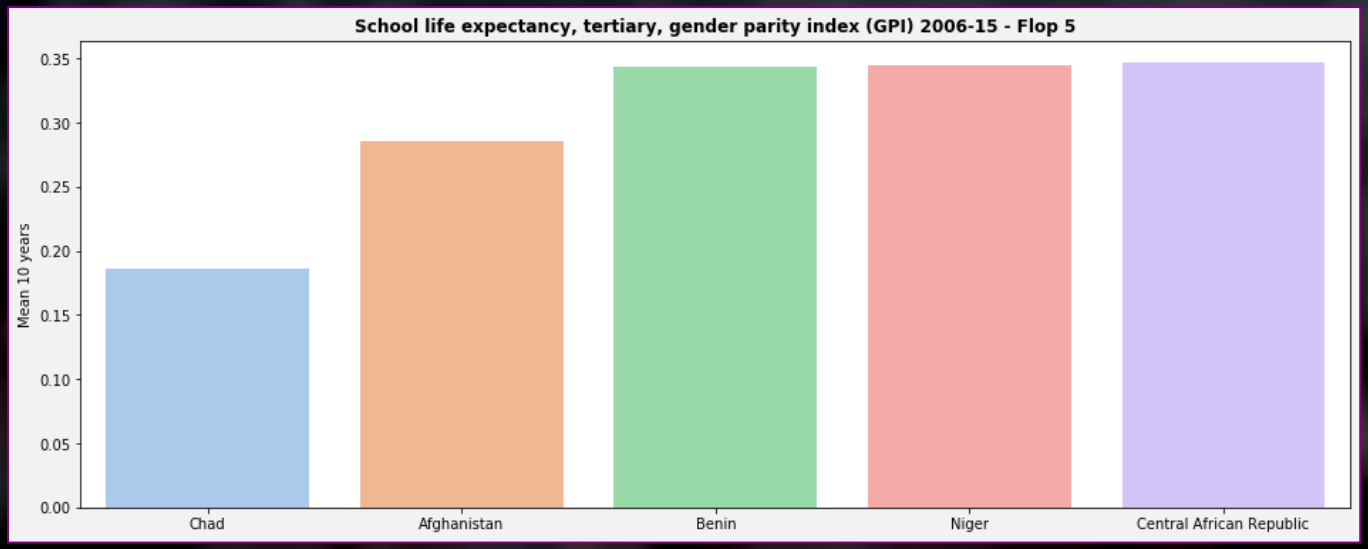


- GPI :
  - > 1 : domination féminine
  - = 1 : parité
  - < 1 : domination masculine
- Afrique Sub-Saharienne & Asie du Sud : écoles supérieures majoritairement terminées par des hommes

ÉTUDE DE L'EXPANSION D'ACADEMY

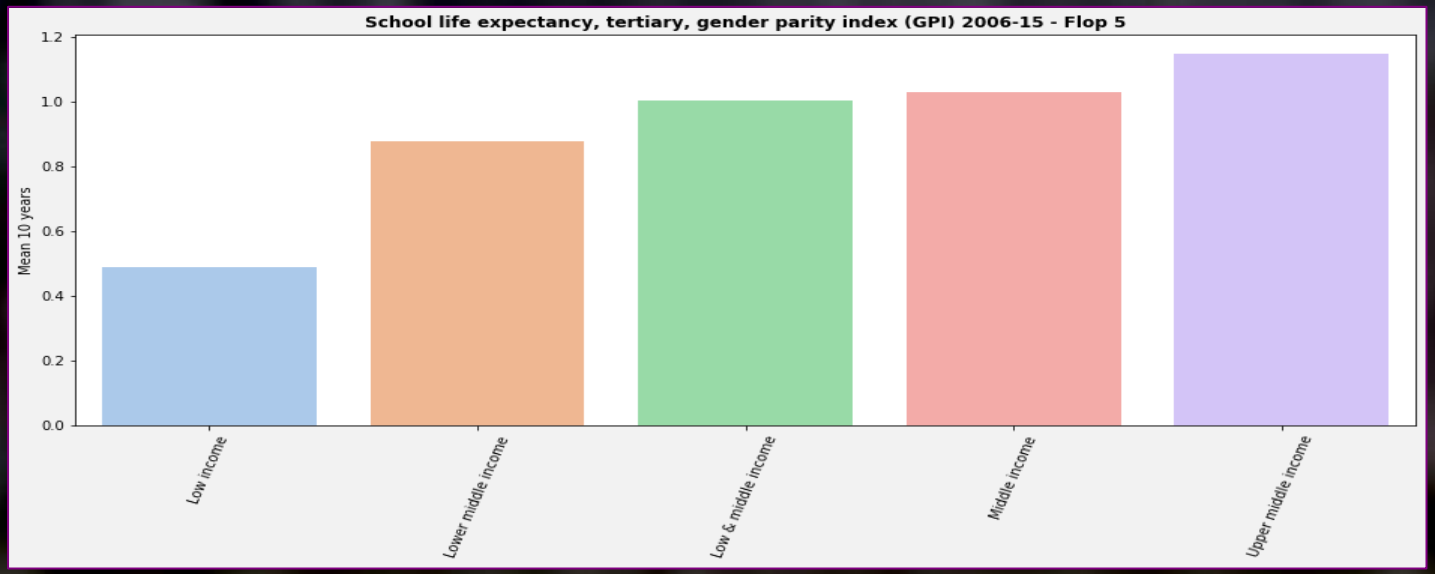
Etudes Tertiaires





- GPI :
  - > 1 : domination féminine
  - = 1 : parité
  - < 1 : domination masculine
- Qatar & Top 5 : écoles supérieures très majoritairement terminées par des femmes
- Chad & Afghanistan : écoles supérieures très majoritairement terminées par des hommes

ÉTUDE DE L'EXPANSION D'ACADEMY



- GPI :
  - > 1 : domination féminine
  - = 1 : parité
  - < 1 : domination masculine
- Pays riches : écoles supérieures terminées légèrement plus par des femmes
- Pays à faibles revenus : écoles supérieures majoritairement terminées par des hommes

CONCLUSION

- Quelques informations à retenir :

- **Pays à moyens revenus :** beaucoup d'inscrits dans le Secondaire (dont privé) et Tertiaire
- **Amérique du Nord :** très connectée, mais peu d'inscrits dans le Secondaire.
- **Asie (Est/Sud) :** beaucoup d'inscrits dans le Secondaire privé et Tertiaire, mais peu de redoublants dans le Secondaire.
- **Chine et Inde :** beaucoup d'enseignants et d'étudiants dans le Secondaire et Tertiaire
- **Programmes :** les plus sollicités dans le Tertiaire : Sciences Sociales / Economie / Droit. Le moins demandé : Agriculture.