OPENCLASSROOMS

Kevin

Parcours Data Scientist

Rappel du sujet/problématique

Entreprise

Start-up de la EdTech : Academy

Propose des contenus de format en ligne pour un public de niveau lycée et université

Problématique

Expansion à l'international

Analyse exploratoire

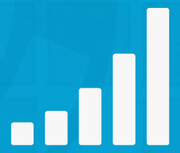
- Quels sont les pays avec un fort potentiel de clients pour nos services ?
- Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise doit-elle opérer en priorité ?

Jeu de données

Analyse pré-exploratoire

Conclusions

1



Jeu de données

5 Fichiers

Banque Mondiale

1. Data
2. Country
3. Country-Series
4. Series
5. FootNote

Fichier Data

Format

886 930 lignes
70 variables

Diversité des données

- Education
- Alphabétisation
- Diplômes
- Dépenses
- Professeurs

Variables

```
----------------------------------------
Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
       '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977', '1978',
       '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986', '1987',
       '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995', '1996',
       '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005',
       '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014',
       '2015', '2016', '2017', '2020', '2025', '2030', '2035', '2040', '2045',
       '2050', '2055', '2060', '2065', '2070', '2075', '2080', '2085', '2090',
       '2095', '2100', 'Unnamed: 69'],
      dtype='object')
```

Pays

242 pays

Indicateurs

3665 indicateurs

Sources variées

- Barro-lee
- DHS
- EGRA
- PISA

5 Fichiers

Banque Mondiale

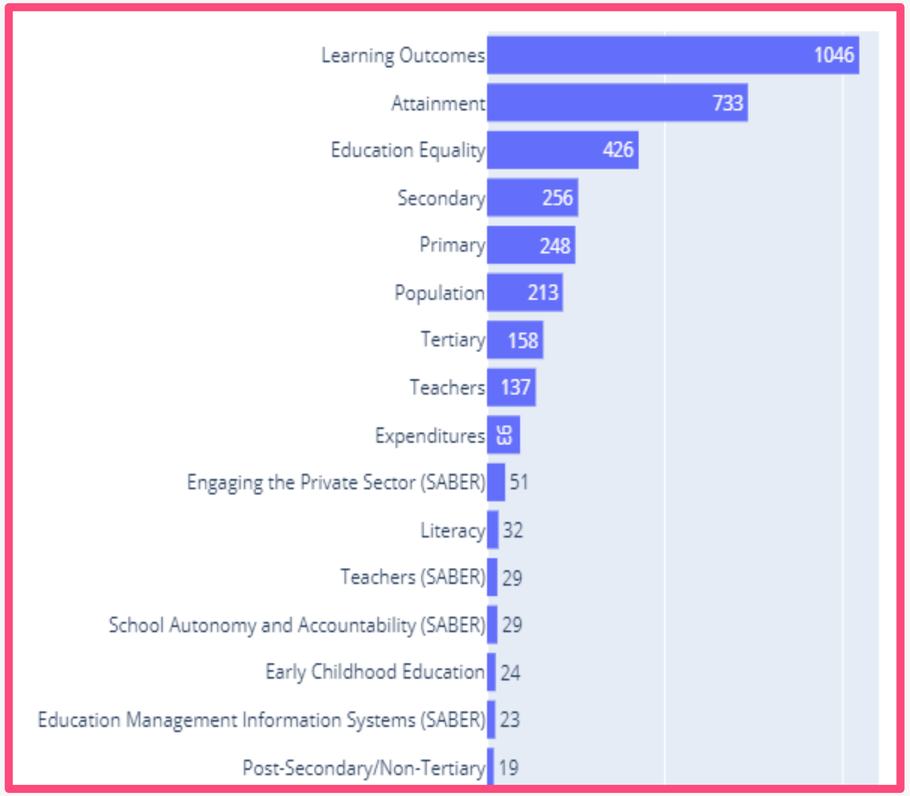
1. Data
2. Country
3. Series
4. Country-Series
5. FootNote

Pertinence des fichiers

Country

```
----------------------------------------
Index(['Country Code', 'Short Name', 'Table Name', 'Long Name', '2-alpha code',
       'Currency Unit', 'Special Notes', 'Region', 'Income Group', 'WB-2 code',
       'National accounts base year', 'National accounts reference year',
       'SNA price valuation', 'Lending category', 'Other groups',
       'System of National Accounts', 'Alternative conversion factor',
       'PPP survey year', 'Balance of Payments Manual in use',
       'External debt Reporting status', 'System of trade',
       'Government Accounting concept', 'IMF data dissemination standard',
       'Latest population census', 'Latest household survey',
       'Source of most recent Income and expenditure data',
       'Vital registration complete', 'Latest agricultural census',
       'Latest industrial data', 'Latest trade data',
       'Latest water withdrawal data', 'Unnamed: 31'],
      dtype='object')
----------------------------------------
```

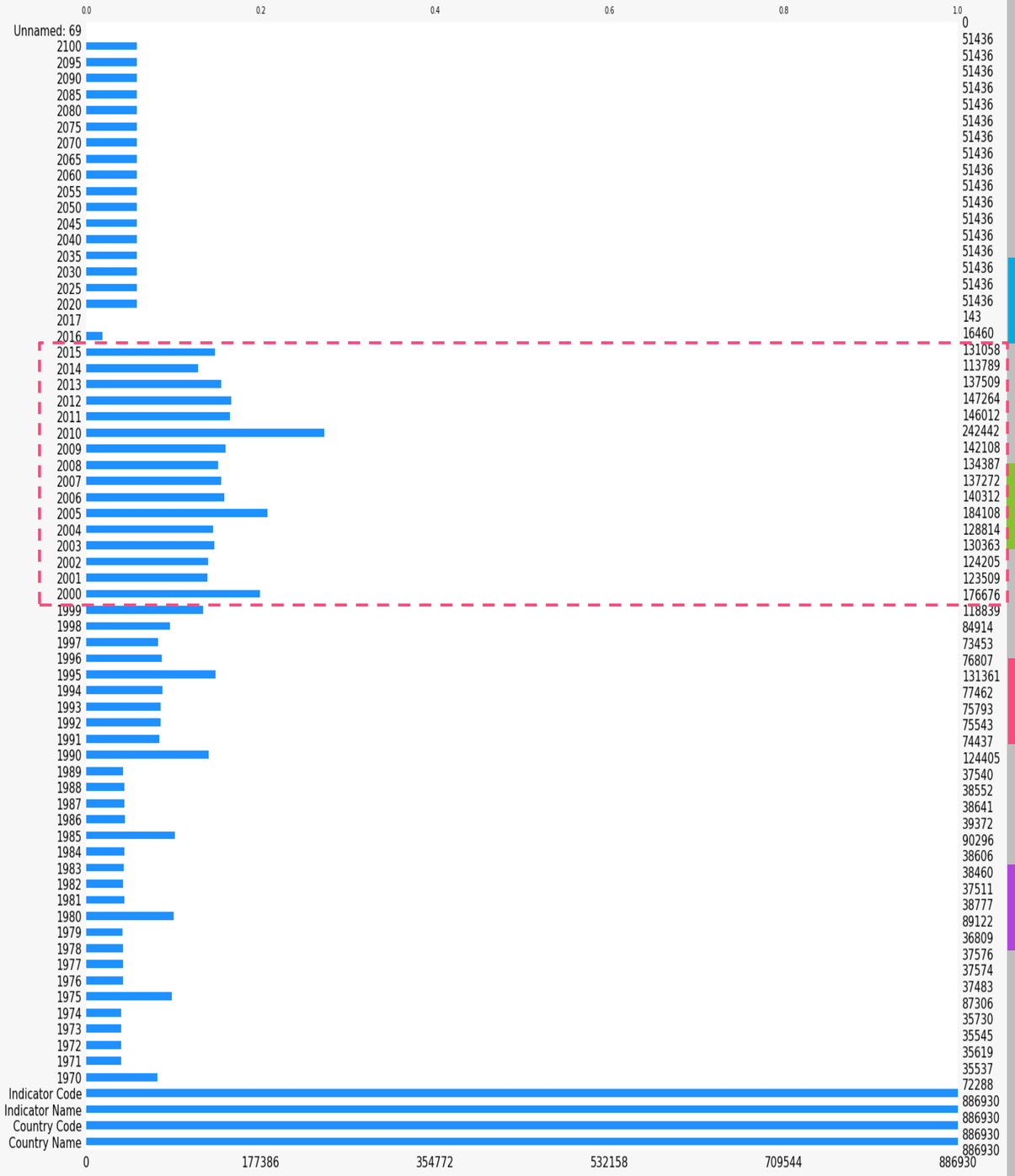
Stats_Series



2

Analyse

Pré-exploratoire



Les années

Réflexion…

1 **Mon expérience**

Excel

2 **Valeurs manquantes**

3 **Pertinence des indicateurs ?**

Hypothèse

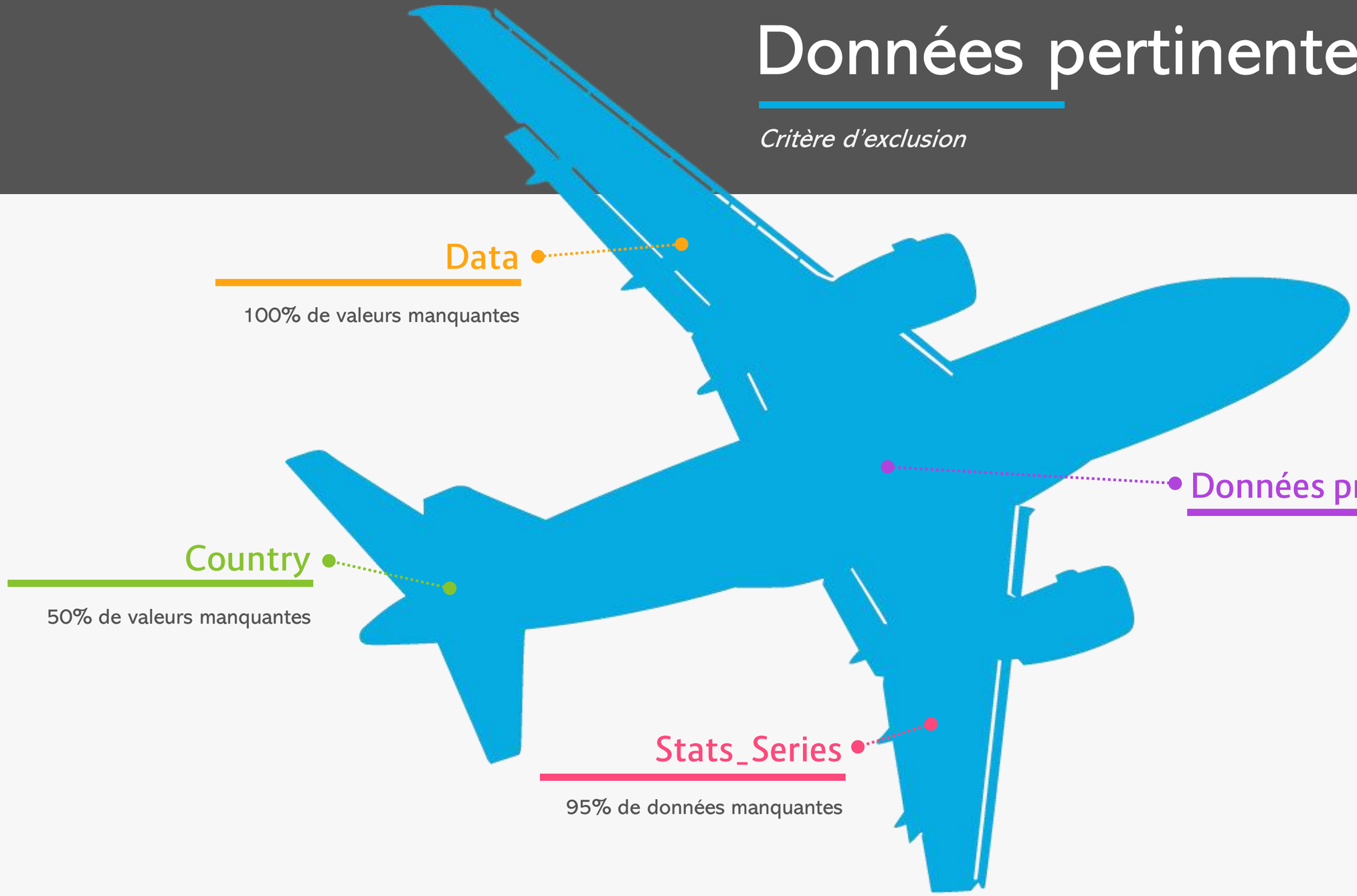
4 **Années à retenir**

2000-2015

Données récentes et plus nombreuses

Données pertinentes

Critère d'exclusion



Data

100% de valeurs manquantes

Country

50% de valeurs manquantes

Stats_Series

95% de données manquantes

Données prêtes à l'emploi

Choix des indicateurs

Accessibilité

- Internet
- Computer

Age

0-14

15-19

20-24

Scolarité

High-School

University

Capacités

- Literacy

Caractéristiques

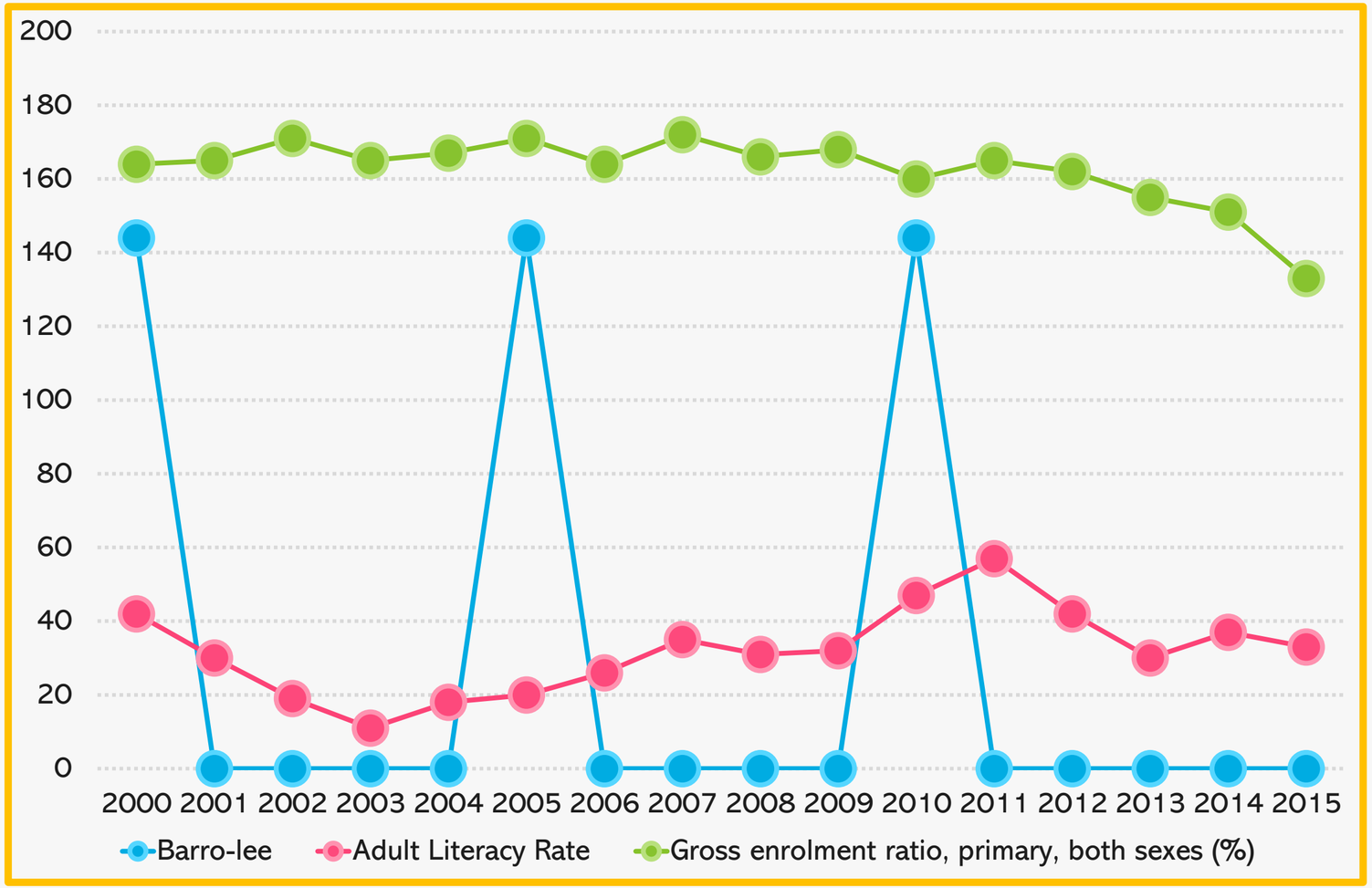
- Total enrollment
- Population growth

Exclusion

Critères divisant la population :

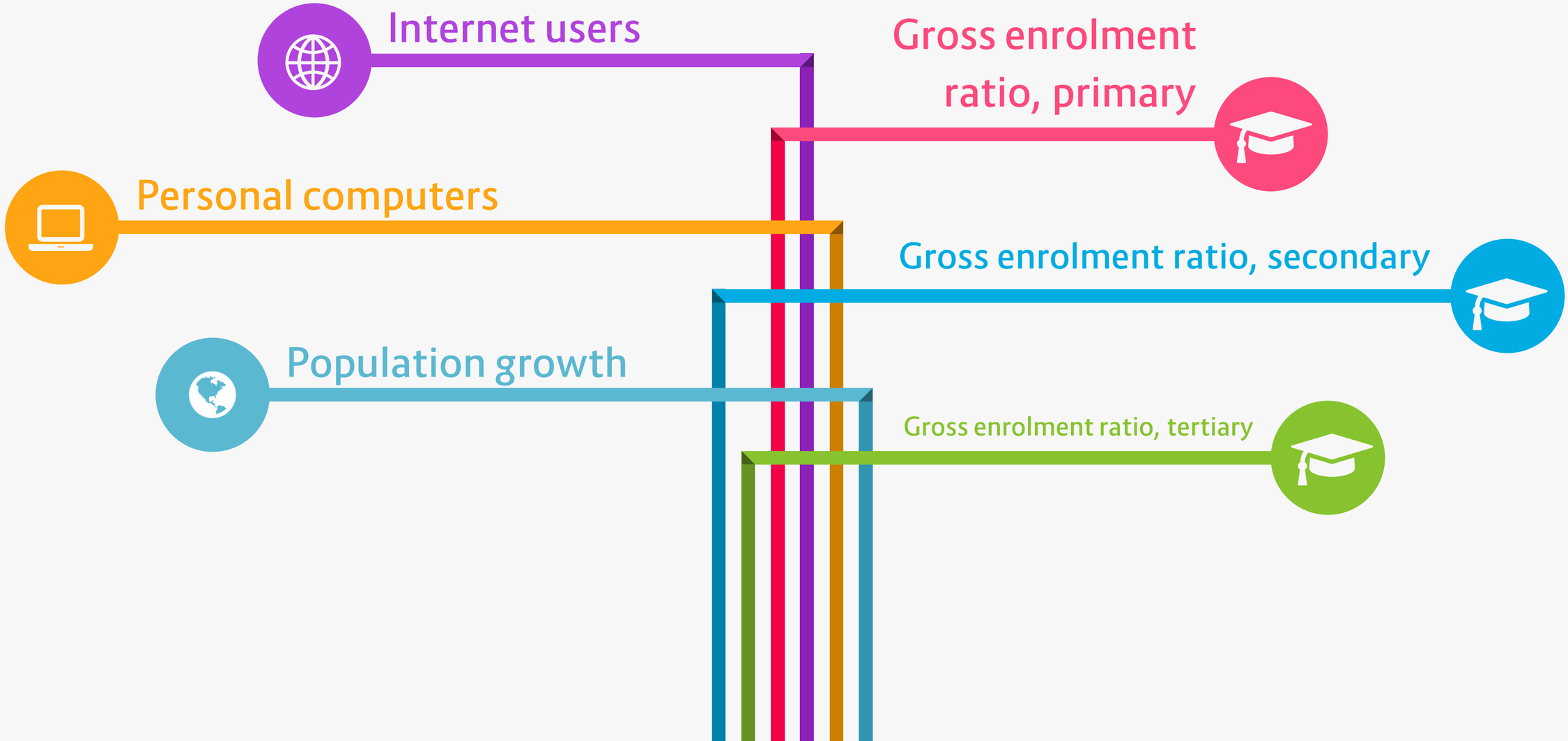
- Sexe
- Level
- Quantile

Vérification hypothèse

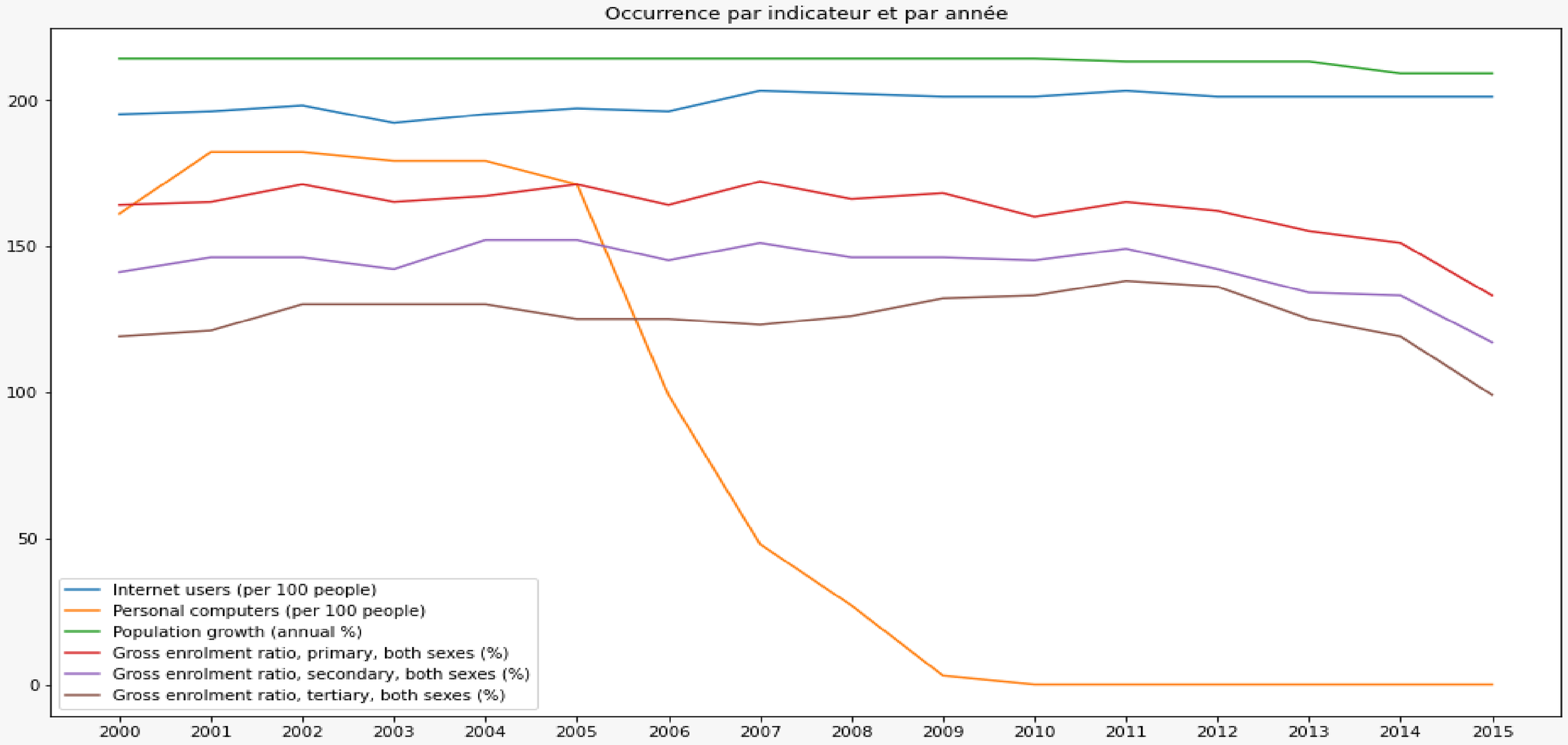


Nombre d’occurences

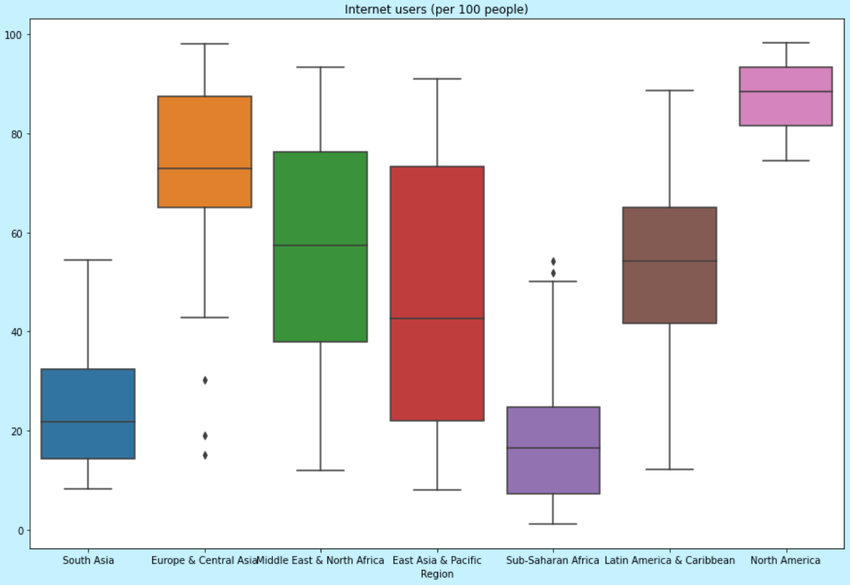
Indicateurs finaux

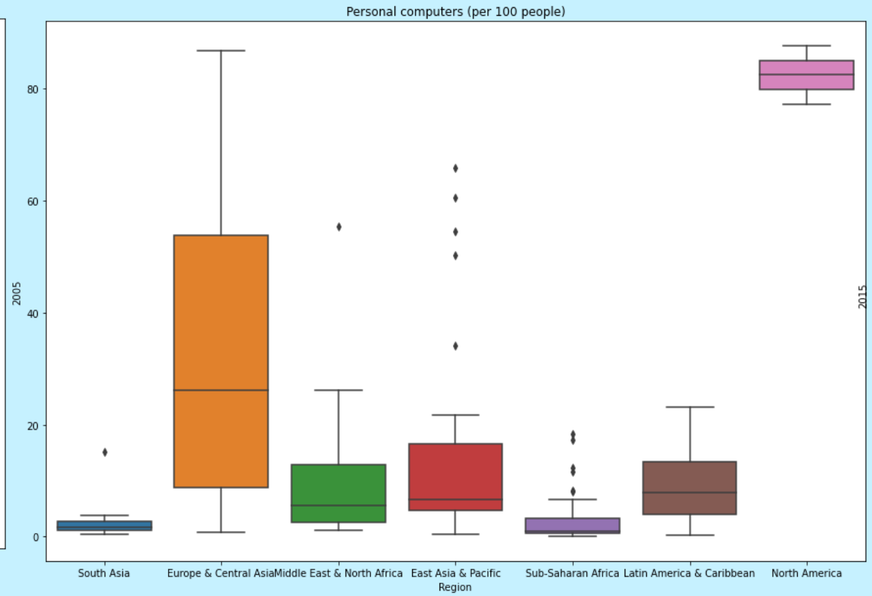


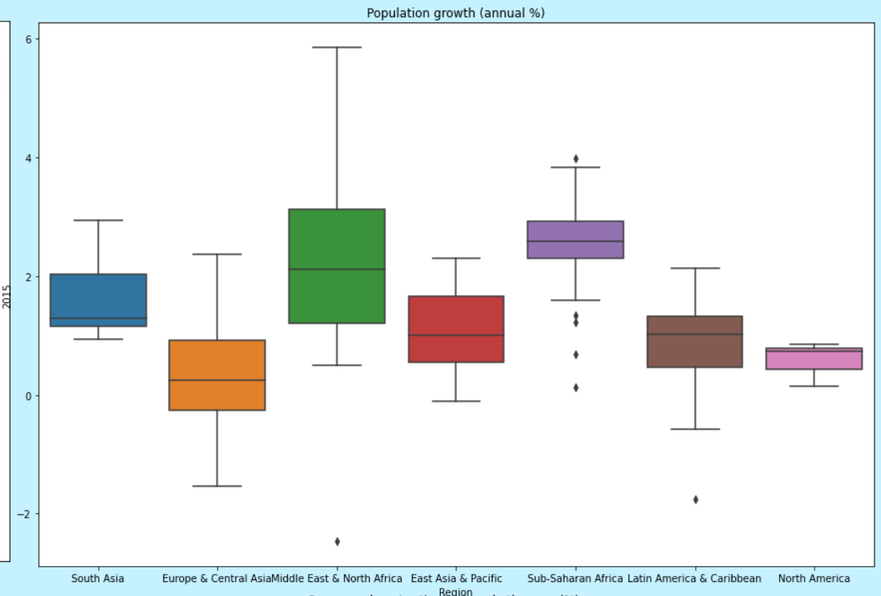
Etude des indicateurs

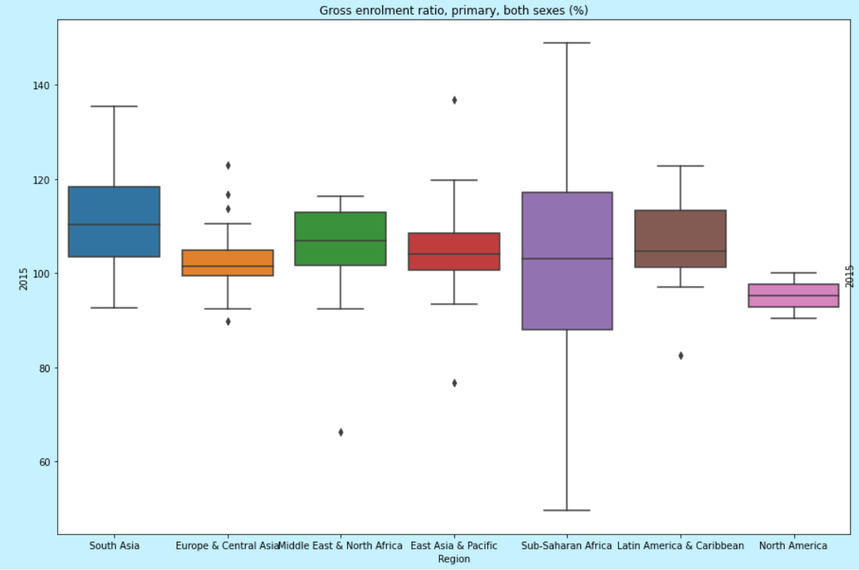


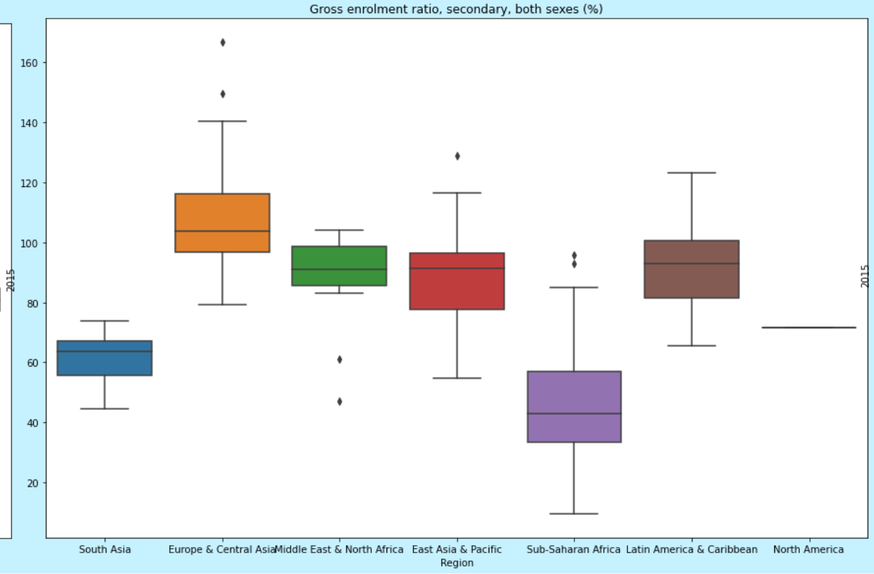
Etude des indicateurs

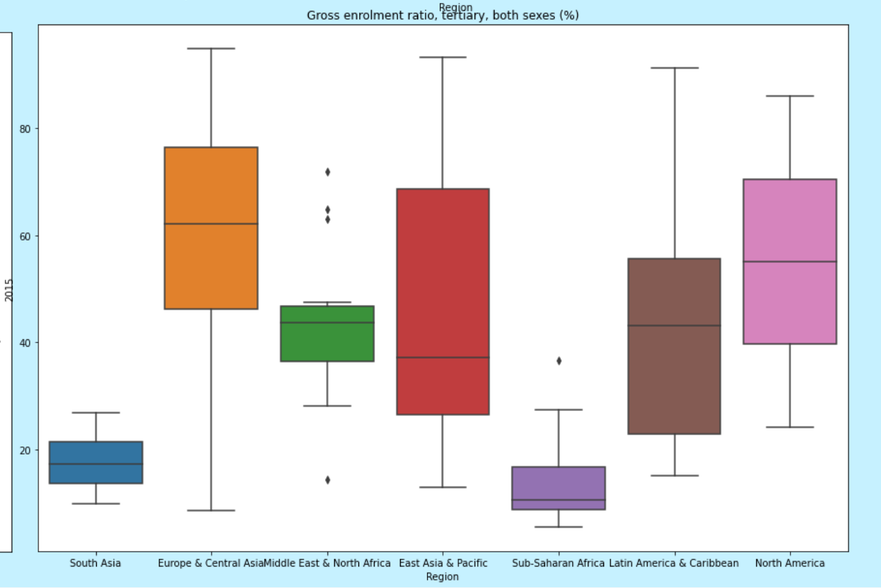












Etude des indicateurs

Corrélation

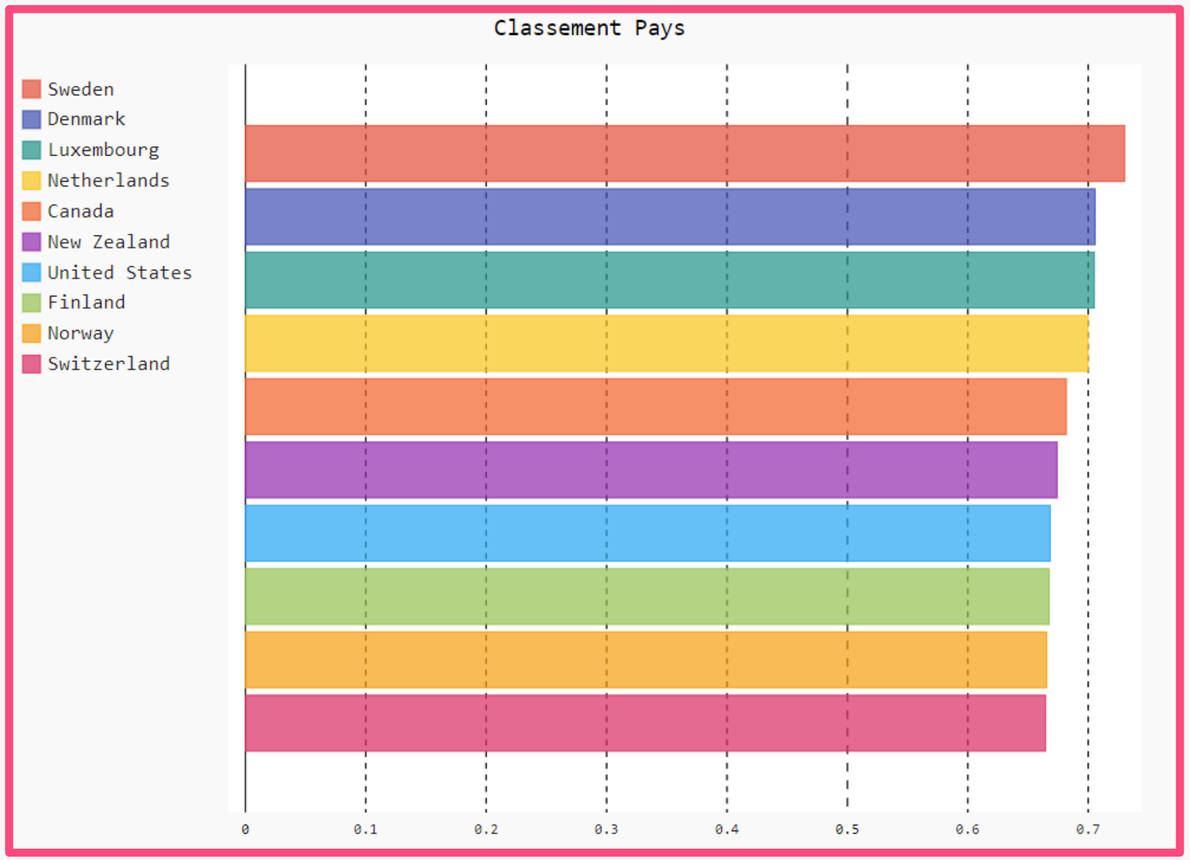
Observation

Scoring

	Region	Income Group	Country Name	internet_users	personal_computers	population_growth	ratio_primary	ratio_secondary	ratio_tertiary
0	South Asia	Low income	Afghanistan	0.084008	0.003897	0.502587	0.751385	0.333583	NaN
1	Europe & Central Asia	Upper middle income	Albania	0.643314	0.019259	-0.049726	0.763627	0.574104	0.613412
2	Middle East & North Africa	Upper middle income	Algeria	0.388513	0.011748	0.327852	0.780113	NaN	0.389754
3	East Asia & Pacific	Upper middle income	American Samoa	NaN	NaN	0.030775	NaN	NaN	NaN
4	Europe & Central Asia	High income: nonOECD	Andorra	0.985623	NaN	-0.262601	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
209	Latin America & Caribbean	High income: nonOECD	Virgin Islands (U.S.)	0.557741	0.031732	-0.097980	NaN	NaN	NaN
210	Middle East & North Africa	Lower middle income	West Bank and Gaza	0.584033	0.062440	0.499420	0.633530	0.497554	0.467456
211	Middle East & North Africa	Lower middle income	Yemen, Rep.	0.244961	0.022640	0.430359	NaN	NaN	NaN
212	Sub-Saharan Africa	Lower middle income	Zambia	0.213580	0.013019	0.516399	NaN	NaN	NaN
213	Sub-Saharan Africa	Low income	Zimbabwe	0.231306	0.076228	0.400542	NaN	NaN	0.089022

Scoring

Pays	Score
Sweden	0.730136
Denmark	0.705662
Luxembourg	0.704975
Netherlands	0.699767
Canada	0.681743
New Zealand	0.674132
United States	0.66839
Finland	0.667478
Norway	0.665429
Switzerland	0.664479

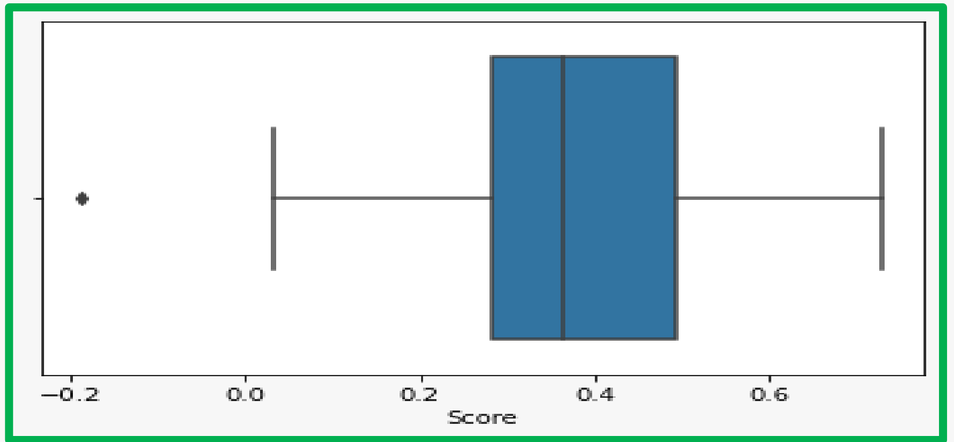


Top 10

- 7 pays européens ou d'Asie Centrale
- 2 pays d'Amerique du Nord
- 1 pays d'Asie Est et Pacifique

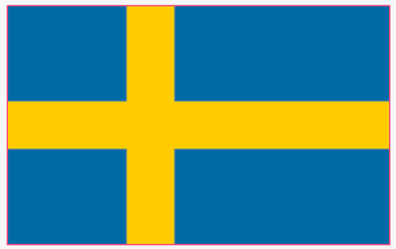
Observations

- Tous les pays ont des revenus élevés
- La Moyenne au Scoring est de 0,38



4

Conclusion

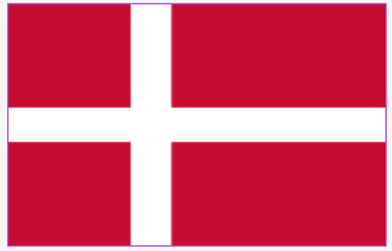


Suède

Finlande

Pays-bas

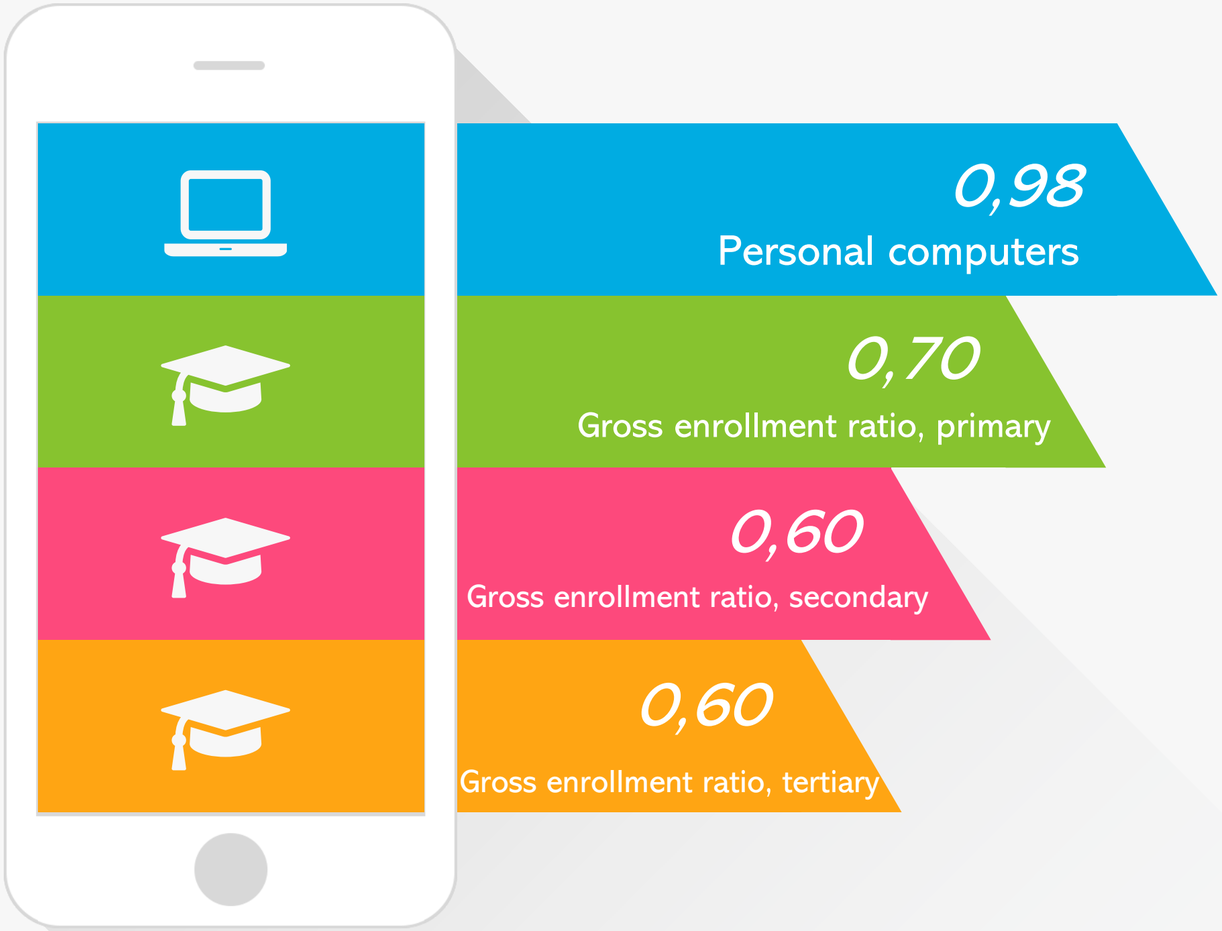
Danemark



Norvège

Scoring

Suisse



0,98
Personal computers

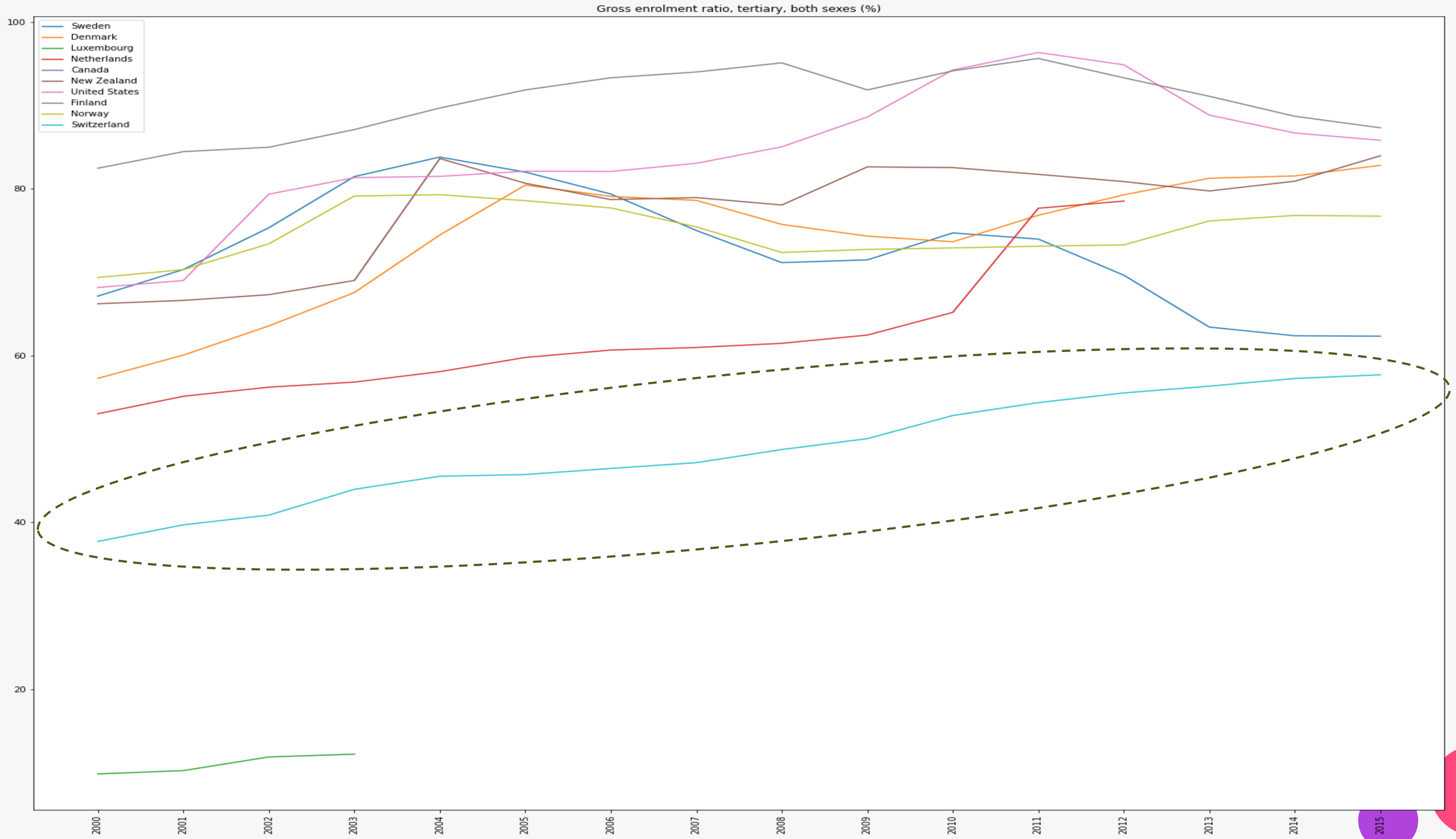
0,70
Gross enrollment ratio, primary

0,60
Gross enrollment ratio, secondary

0,60
Gross enrollment ratio, tertiary

Scoring : 10ème

Mais des défauts...


Gross enrolment ratio, tertiary, both sexes (%)
Sweden
Denmark
Luxembourg
Netherlands
Canada
New Zealand
United States
Finland
Norway
Switzerland
100
80
60
40
20
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Scoring Région

	Region	Income Group	Country Name	internet_users	personal_computers	population_growth	ratio_primary	ratio_secondary	ratio_tertiary	Score
184	Europe & Central Asia	High income: OECD	Sweden	0.921551	0.952792	0.180571	0.826217	0.842037	0.657651	0.730136
52	Europe & Central Asia	High income: OECD	Denmark	0.979729	0.793632	0.120629	0.681798	0.784281	0.873904	0.705662
113	Europe & Central Asia	High income: OECD	Luxembourg	0.989936	0.722003	0.402985	NaN	NaN	NaN	0.704975
136	Europe & Central Asia	High income: OECD	Netherlands	0.932880	0.974924	0.075684	0.703194	0.812152	NaN	0.699767
34	North America	High income: OECD	Canada	0.899784	1.000000	0.145446	NaN	NaN	NaN	0.681743
138	East Asia & Pacific	High income: OECD	New Zealand	0.897271	0.572684	0.322573	0.667324	0.698955	0.885984	0.674132
203	North America	High income: OECD	United States	0.758253	0.880818	0.124609	0.672604	NaN	0.905666	0.668390
65	Europe & Central Asia	High income: OECD	Finland	0.878956	0.570338	0.056246	0.681917	0.895974	0.921435	0.667478
143	Europe & Central Asia	High income: OECD	Norway	0.984609	0.676628	0.169921	0.674444	0.677359	0.809612	0.665429
185	Europe & Central Asia	High income: OECD	Switzerland	0.889705	0.989318	0.194383	0.697911	0.606768	0.608791	0.664479

Observations

- L'Amerique du Nord est première en terme de score par Région
- Le Canada est première en terme d'ordinateur personnel

A suivre

- Faible échantillon en Amerique du nord

Region	internet_users	personal_computers	population_growth	ratio_primary	ratio_secondary	ratio_tertiary	Score
North America	0.886012	0.940409	0.098407	0.639700	0.428890	0.581093	0.595752
Europe & Central Asia	0.737239	0.366868	0.060891	0.687810	0.651669	0.642272	0.524458
Middle East & North Africa	0.563925	0.126780	0.387037	0.697724	0.522337	0.461816	0.459937
East Asia & Pacific	0.481339	0.178317	0.191893	0.706181	0.536528	0.490151	0.430735
Latin America & Caribbean	0.540023	0.096465	0.152345	0.710789	0.551373	0.472867	0.420644
South Asia	0.259974	0.039867	0.273085	0.750790	0.365490	0.188792	0.313000
Sub-Saharan Africa	0.184971	0.032419	0.433721	0.690604	0.284525	0.151371	0.296268

Limites

Indicateurs

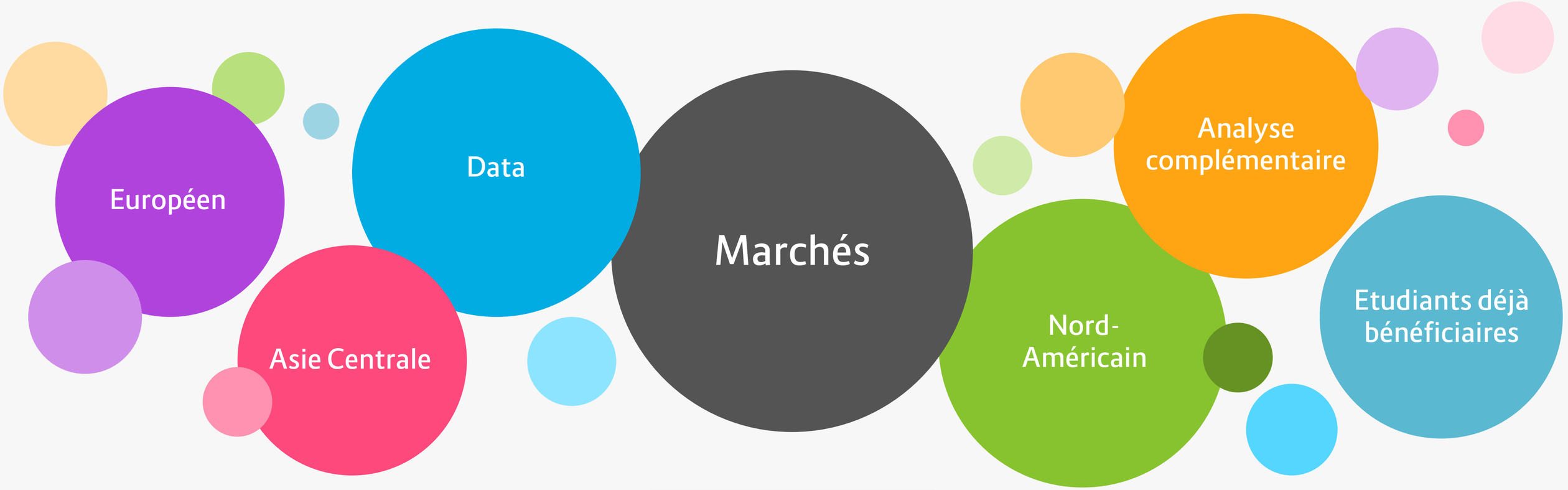
Manques de données

Combinaisons possibles

Stratégie

Explications

Conclusion


Européen
Data
Asie Centrale
Marchés
Nord-
Américain
Analyse
complémentaire
Etudiants déjà
bénéficiaires