



# **Parcours Data Scientist**

## **Projet N°2 : Analysez des données de systèmes éducatifs**

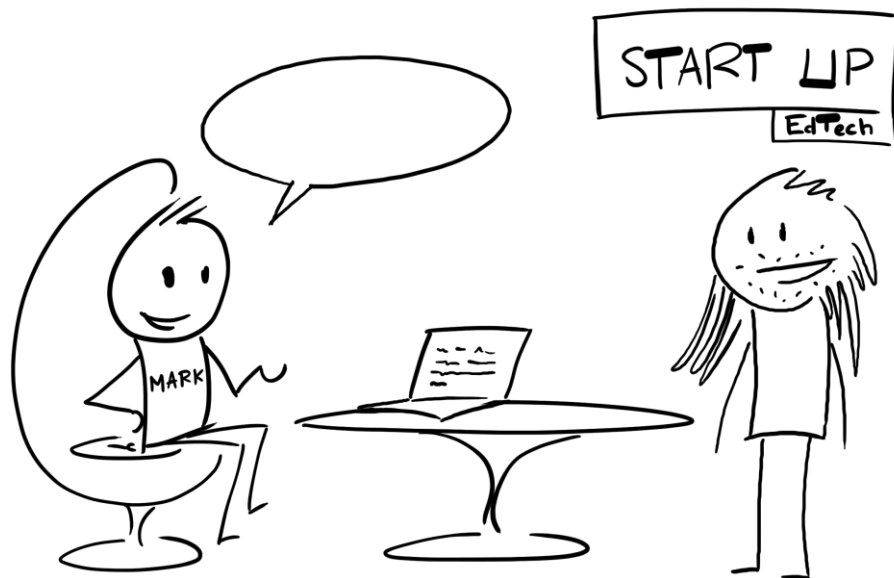
**Daniel CHASTANET**



- **Rappel de la problématique**
- **Présentation du jeu de données**
- **Analyse pré-exploratoire**
- **Analyse**
- **Classement**
- **Conclusion**



## Rappel de la problématique



## Projet d'expansion à l'international de l'entreprise Première mission d'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 ?**








## Présentation du jeu de données



Les données que l'on a à notre disposition sont celle de banque mondiale

<https://datacatalog.worldbank.org/dataset/education-statistics>

**5 bases de données :**

-  EdStatsCountry.csv
-  EdStatsCountry-Series.csv
-  EdStatsData.csv
-  EdStatsFootNote.csv
-  EdStatsSeries.csv



## Présentation du jeu de données – Country (data\_1)



### Descriptif des paramètres qui définissent les bases de données : Noms des pays / Revenus / Années références des données

dimension : (241, 32)

nombre de données dupliquées : 0

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 241 entries, 0 to 240  
Columns: 32 entries, Country Code to Unnamed: 31  
dtypes: float64(4), object(28)  
memory usage: 60.4+ KB
```

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD	AW	...
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income	AF	...

% de données manquantes

Country Code	0.000000
Short Name	0.000000
Table Name	0.000000
Long Name	0.000000
2-alpha code	1.244813
Currency Unit	10.788382
Special Notes	39.834025
Region	11.203320
Income Group	11.203320
WB-2 code	0.414938
National accounts base year	14.937759
National accounts reference year	86.721992
SNA price valuation	18.257261
Lending category	40.248963
Other groups	75.933610
System of National Accounts	10.788382
Alternative conversion factor	80.497925
PPP survey year	39.834025
Balance of Payments Manual in use	24.896266
External debt Reporting status	48.547718
System of trade	17.012448
Government Accounting concept	33.195021
IMF data dissemination standard	24.896266
Latest population census	11.618257
Latest household survey	41.493776
Source of most recent Income and expenditure data	33.609959
Vital registration complete	53.941909
Latest agricultural census	41.078838
Latest industrial data	55.601660
Latest trade data	23.236515
Latest water withdrawal data	25.726141
Unnamed: 31	100.000000
dtype: float64	



# Présentation du jeu de données – Series (data\_5)



dimension : (3665, 21)

nombre de données dupliquées : 0

% de données manquantes

Series Code	0.000000
Topic	0.000000
Indicator Name	0.000000
Short definition	41.173261
Long definition	0.000000
Unit of measure	100.000000
Periodicity	97.298772
Base Period	91.432469
Other notes	84.938608
Aggregation method	98.717599
Limitations and exceptions	99.618008
Notes from original source	100.000000
General comments	99.618008
Source	0.000000
Statistical concept and methodology	99.372442
Development relevance	99.918145
Related source links	94.133697
Other web links	100.000000
Related indicators	100.000000
License Type	100.000000
Unnamed: 20	100.000000
dtype: float64	

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age 15-19 with no education	Percentage of female population age 15-19 with no education	Percentage of female population age 15-19 with no education	NaN	NaN	NaN	NaN	NaN	...
1	BAR.NOED.1519.ZS	Attainment	Barro-Lee: Percentage of population age 15-19 with no education	Percentage of population age 15-19 with no education	Percentage of population age 15-19 with no education	NaN	NaN	NaN	NaN	NaN	...

## Description des indicateurs, réunis en catégorie, et des méthode de prise de données

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Columns: 21 entries, Series Code to Unnamed: 20
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```



# Présentation du jeu de données – Data (data\_3/Country-S)



dimension : (886930, 70)

nombre de données dupliquées : 0

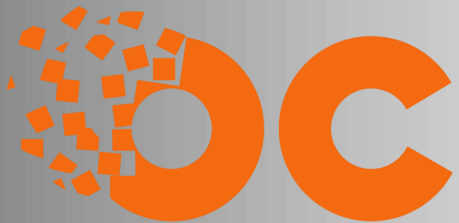
% de données manquantes

Country Name	0.000000
Country Code	0.000000
Indicator Name	0.000000
Indicator Code	0.000000
1970	91.849639
...	
2085	94.200670
2090	94.200670
2095	94.200670
2100	94.200670
Unnamed: 69	100.000000
Length: 70, dtype: float64	

	Country Name	Country Code	Indicator Name	Indicator Code	1970	...	2085	2090	2095	2100
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	...	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	...	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	...	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	...	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	...	NaN	NaN	NaN	NaN

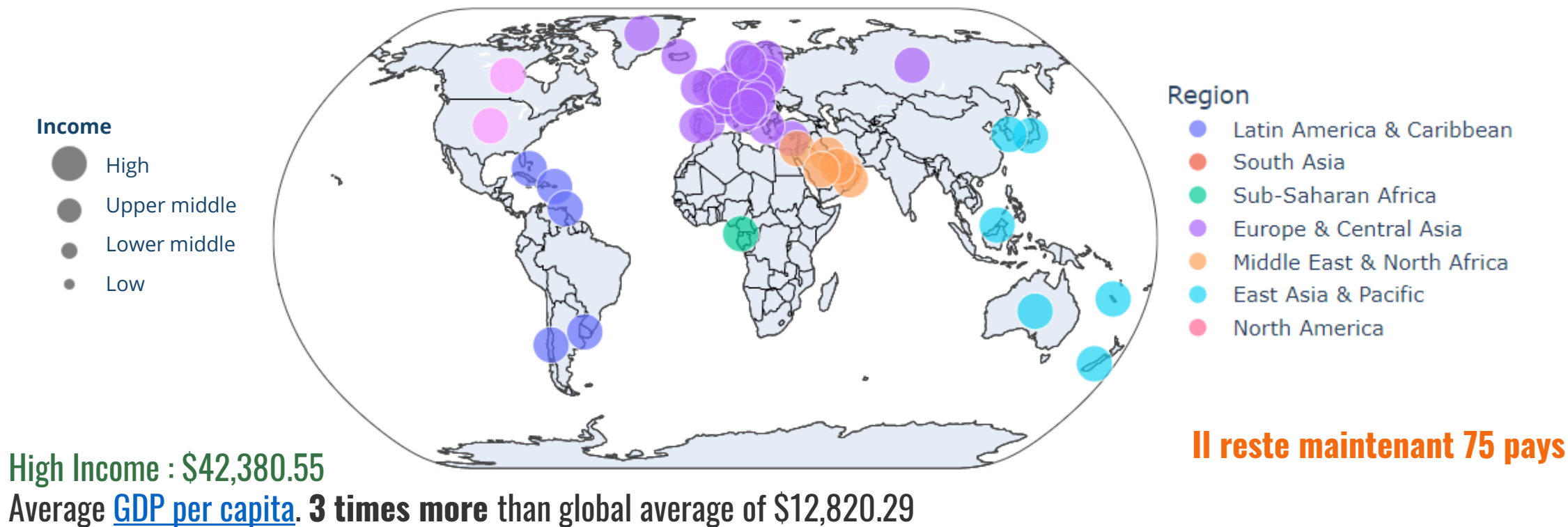
**Données numériques par années (1970 - ... - 2100) en fonction des différents indicateurs de la banque mondiale**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Columns: 70 entries, Country Name to Unnamed: 69
dtypes: float64(66), object(4)
memory usage: 473.7+ MB
```



## Analyse pré-exploratoire (data\_1/Country)

D'après data\_1, on peut déjà faire un tri sur le groupe d'appartenance en terme de revenu des pays







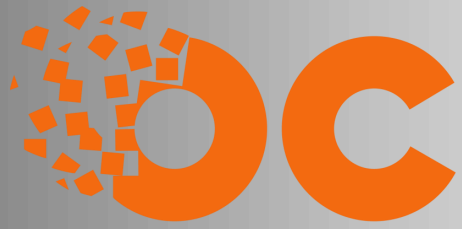
## Analyse pré-exploratoire (data\_5/Series)

### Liste des catégories

```
'Education Equality',  
'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',  
'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)',  
'Background',  
'Primary',  
'Secondary',  
'Tertiary',  
'Early Childhood Education',  
'Pre-Primary',  
'Health: Risk factors',  
'Health: Mortality',  
'Social Protection & Labor: Labor force structure',  
'Labor',  
'Social Protection & Labor: Unemployment',  
'Health: Population: Structure',  
'EMIS',  
'Post-Secondary/Non-Tertiary']
```

### Catégories sélectionnées :

- **Attainment**
- **Infrastructure : Communications**
- **Expenditures**
- **Population**
- **Health: Population: Dynamics**
- **Health: Population: Structure**



## Analyse pré-exploratoire (data\_5/ Series)

**Pour chaque catégorie, je recherche l'indicateur le plus pertinent pour la problématique :**

**Attainment : UIS.EA.2.AG25T99 → Percentage of population age 25+ with completed lower secondary education**

**Infrastructure : Communications : IT.NET.USER.P2 → Internet users (per 100 people)**

**Expenditures : UIS.XUNIT.US.56.FSGOV → Government expenditure per tertiary student (US\$)**  
**UIS.XUNIT.US.23.FSGOV → Government expenditure per secondary student (US\$)**

**Population : SP.POP.1524.TO.UN → Population, ages 15-24**

**Health: Population: Dynamics : SP.POP.GROW → Population growth (annual %)**

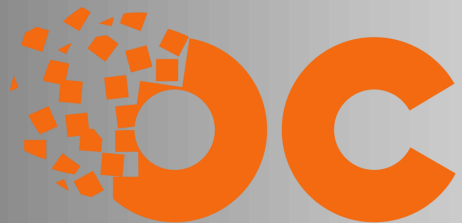
**Health: Population: Structure : SP.POP.1564.TO → Population ages 15-64**



## Analyse pré-exploratoire (data\_1 / data\_5 / data\_3)

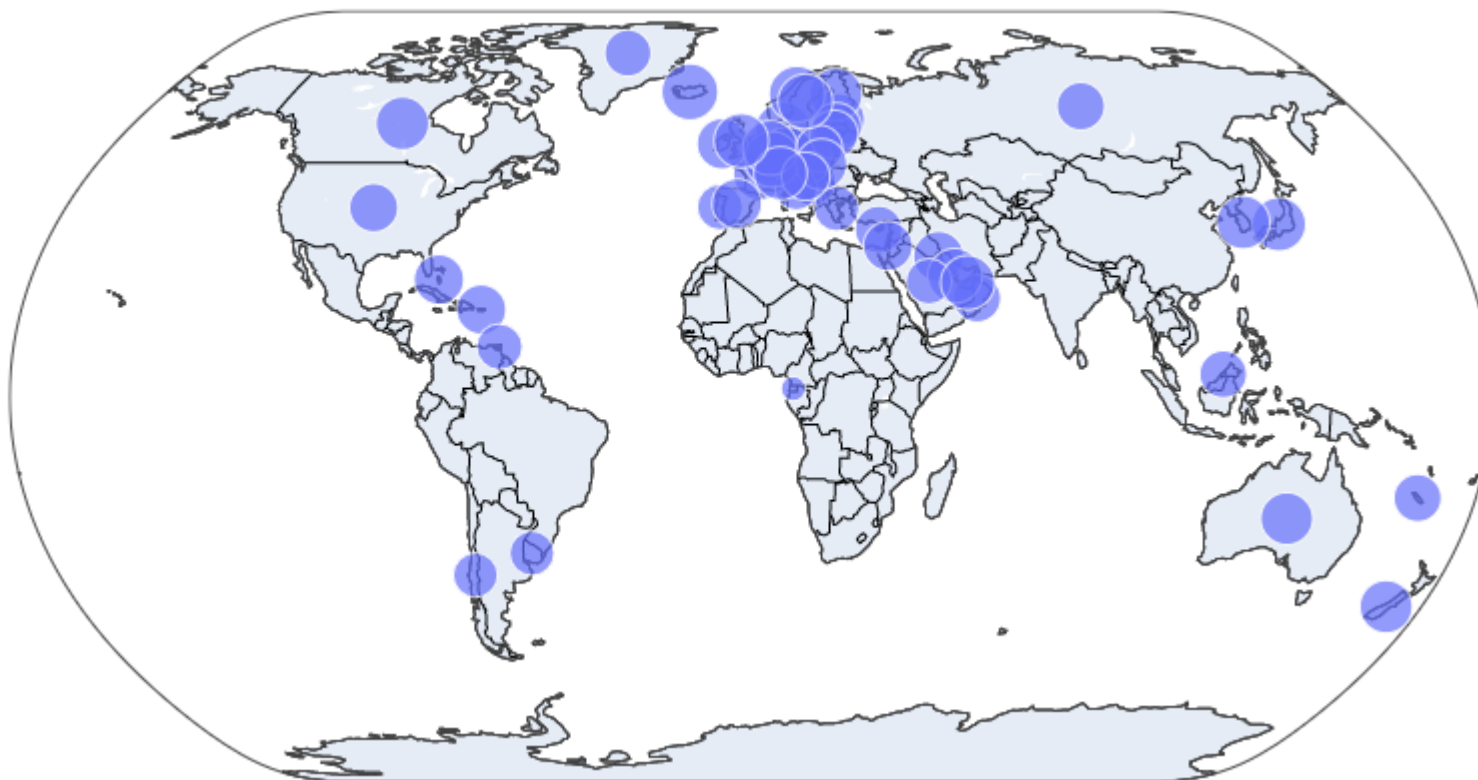
### Jointure ou filtrage des données

	Country Name	Country Code	Indicator Name	Indicator Code	2012	2013	2014	2015	2020	2025	Series Code	Topic
0	Andorra	AND	All staff compensation as % of total expenditure in lower secondary public institutions (%)	UIS.XSPENDP.2.FDPUB.FNS	93.206421	64.498459	66.257217	NaN	NaN	NaN	UIS.XSPENDP.2.FDPUB.FNS	Expenditures
1	Andorra	AND	All staff compensation as % of total expenditure in post-secondary non-tertiary public institutions (%)	UIS.XSPENDP.4.FDPUB.FNS	NaN	NaN	NaN	NaN	NaN	NaN	UIS.XSPENDP.4.FDPUB.FNS	Expenditures
2	Andorra	AND	All staff compensation as % of total expenditure in pre-primary public institutions (%)	UIS.XSPENDP.0.FDPUB.FNS	62.544651	62.563992	62.604301	NaN	NaN	NaN	UIS.XSPENDP.0.FDPUB.FNS	Expenditures

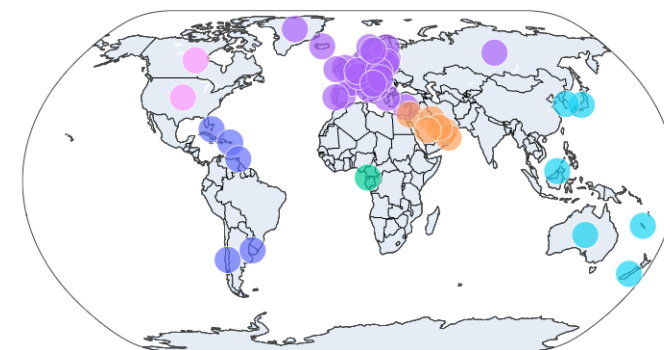


## Analyse exploratoire

Utilisateurs internet pour 100 personnes



Hauts revenus

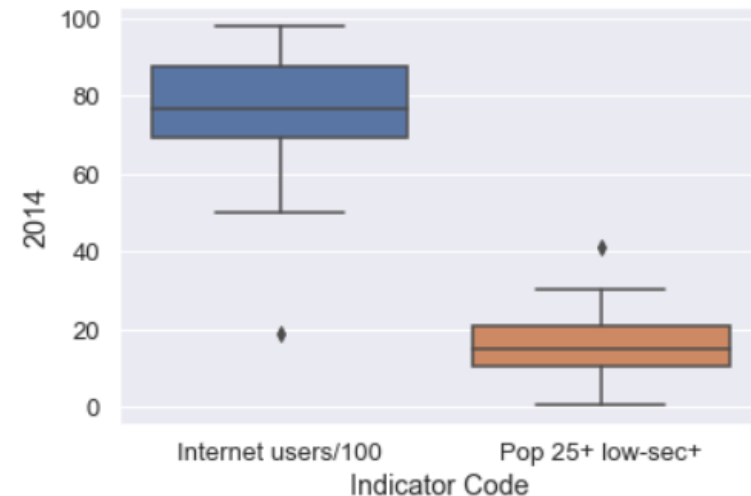
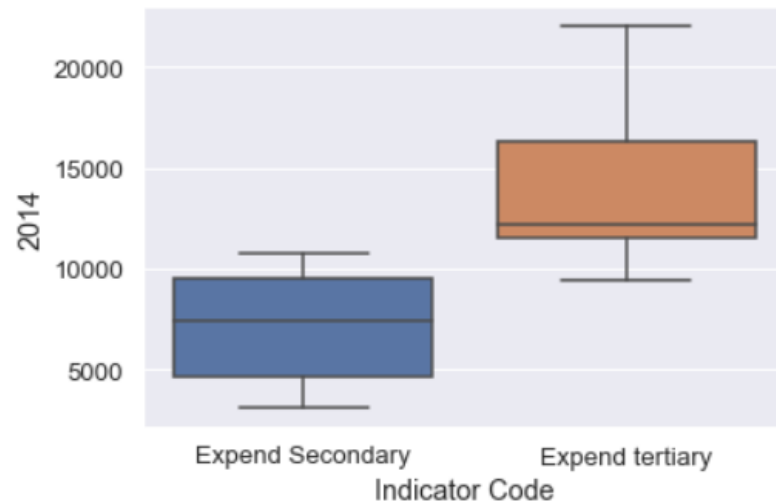
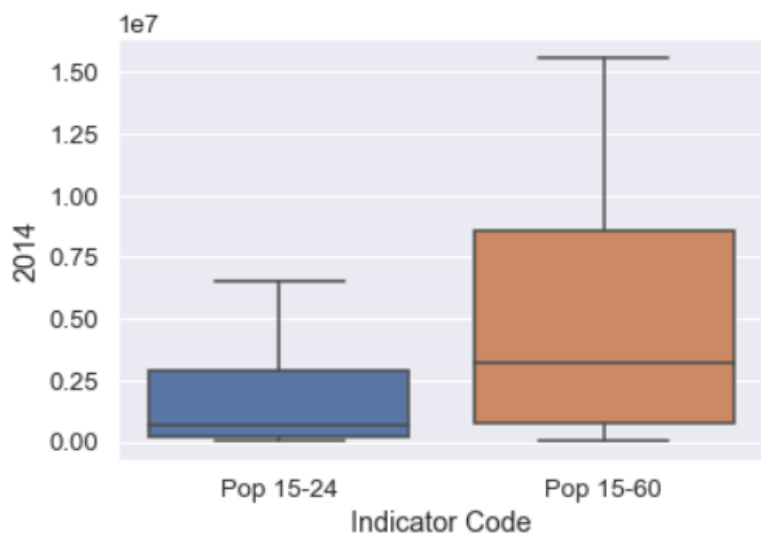


Nombre  
d'utilisateurs  
élevés dans les  
pays à fort revenu

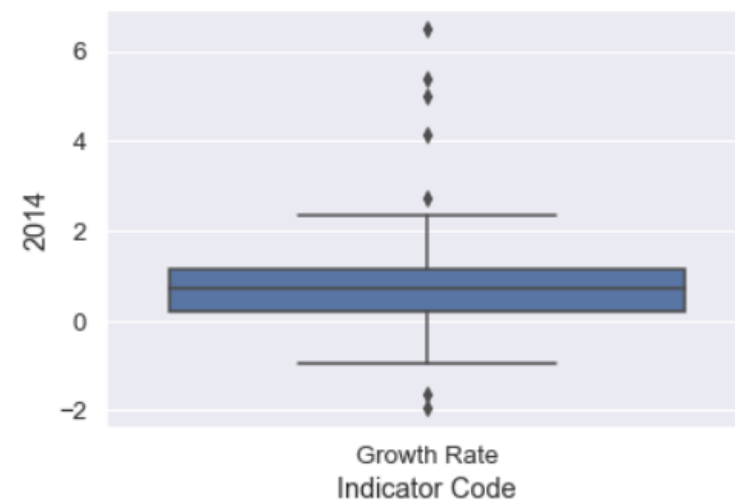
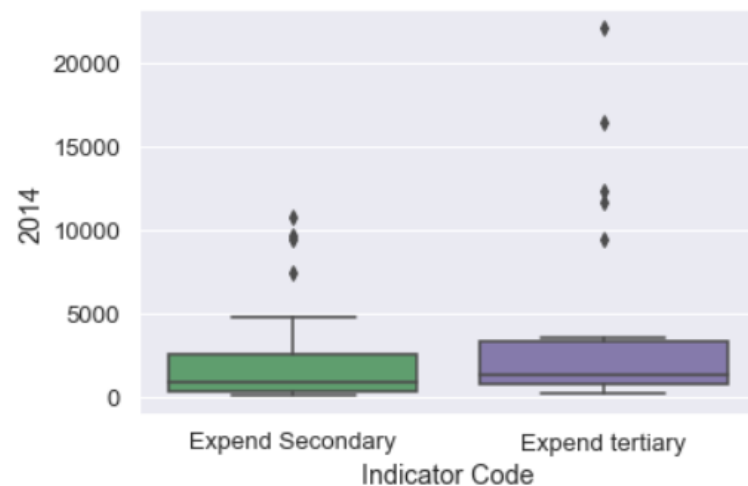


# Analyse exploratoire

## Boîte à moustaches des données des Pays à « hauts revenus » en fonction des indicateurs



## World





## Classement des pays (année de référence 2014)

(525, 12)

-----

Données manquantes (%)

Country Name 0.000000

Country Code 0.000000

Indicator Name 0.000000

Indicator Code 0.000000

2012 34.857143

2013 38.857143

2014 43.428571

2015 52.000000

2020 100.000000

2025 100.000000

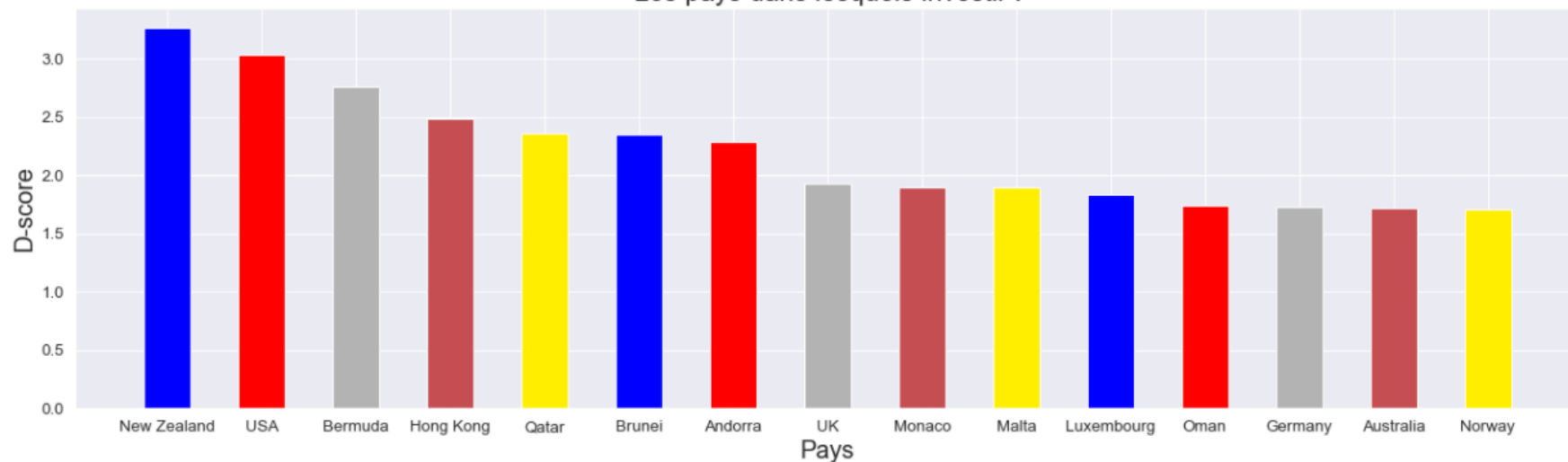
Series Code 0.000000

Topic 0.000000

dtype: float64

	Country Name	Country Code	Indicator Name	Indicator Code	2014	Series Code
0	Andorra	AND	Government expenditure per secondary student (US\$)	UIS.XUNIT.US.23.FSGOV	4769.050293	UIS.XUNIT.US.23.FSGOV
1	Andorra	AND	Government expenditure per tertiary student (US\$)	UIS.XUNIT.US.56.FSGOV	12254.121094	UIS.XUNIT.US.56.FSGOV
2	Andorra	AND	Internet users (per 100 people)	IT.NET.USER.P2	95.900000	IT.NET.USER.P2

Les pays dans lesquels investir !



**Normalisation**  
**Σ des 7 indicateurs**

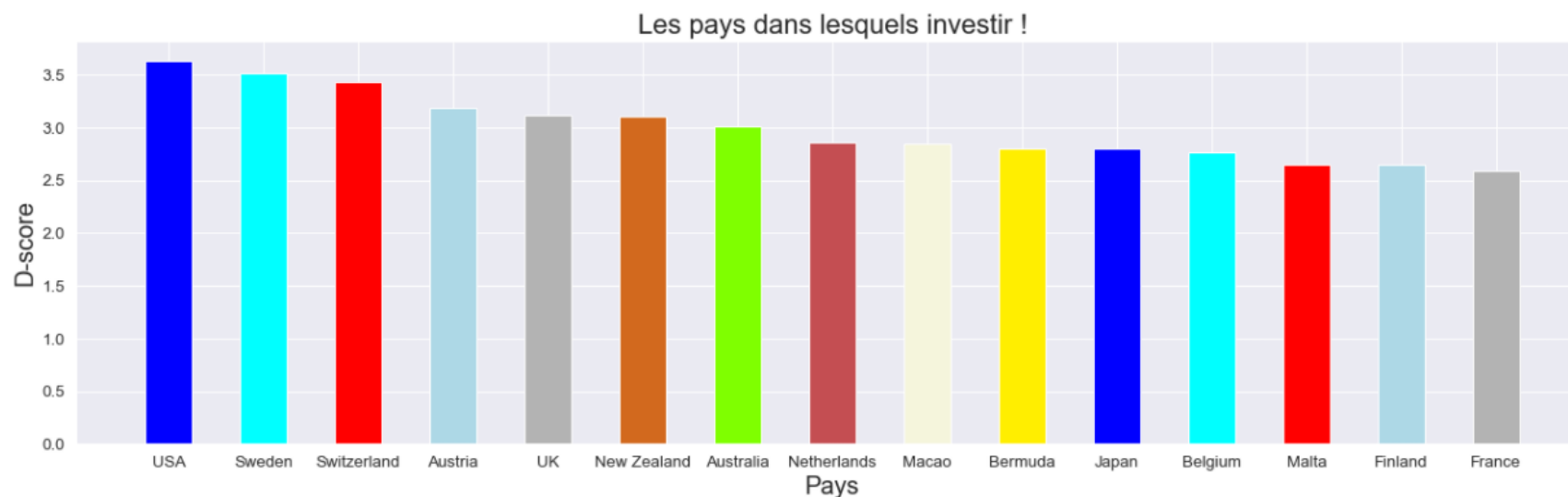


## Classement

	Country Name	Country Code	Indicator Name	Indicator Code	2012	2013	2014	2015	2020	2025	Series Code	Topic
0	Andorra	AND	Government expenditure per secondary student (US\$)	UIS.XUNIT.US.23.FSGOV	NaN	6652.989746	4769.050293	NaN	NaN	NaN	UIS.XUNIT.US.23.FSGOV	Expenditure
1	Andorra	AND	Government expenditure per tertiary student (US\$)	UIS.XUNIT.US.56.FSGOV	NaN	6824.612305	12254.121094	NaN	NaN	NaN	UIS.XUNIT.US.56.FSGOV	Expenditure
2	Andorra	AND	Internet users (per 100 people)	IT.NET.USER.P2	86.434425	94.000000	95.900000	96.910000	NaN	NaN	IT.NET.USER.P2	Infrastructure Communication

ponderation :

- 1 Gov exp secondary \* 1
- 2 Gov exp tertiary \* 1
- 3 Internet user/100 \* 1
- 4 Growth Rate \* 1
- 5 Pop 15-24 \* 1
- 6 Pop 15-60 \* 1
- 7 % Pop 25+ lowsec \* 1





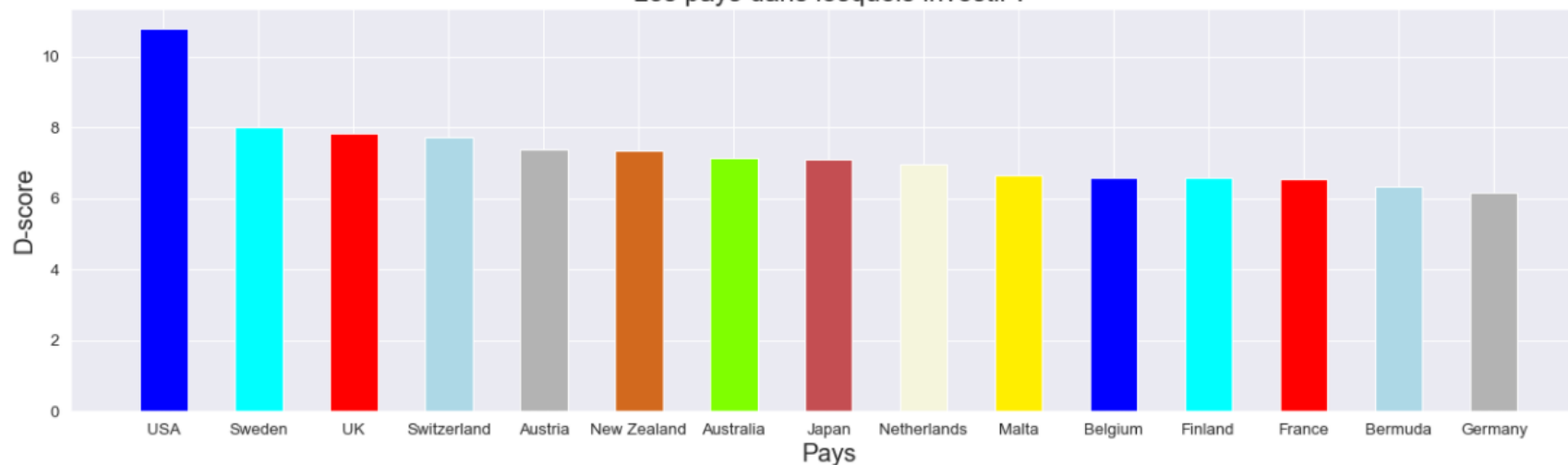
## Classement

	Country Name	Country Code	Indicator Name	Indicator Code	2012	2013	2014	2015	2020	2025	Series Code	Topic
0	Andorra	AND	Government expenditure per secondary student (US\$)	UIS.XUNIT.US.23.FSGOV	NaN	6652.989746	4769.050293	NaN	NaN	NaN	UIS.XUNIT.US.23.FSGOV	Expenditure
1	Andorra	AND	Government expenditure per tertiary student (US\$)	UIS.XUNIT.US.56.FSGOV	NaN	6824.612305	12254.121094	NaN	NaN	NaN	UIS.XUNIT.US.56.FSGOV	Expenditure
2	Andorra	AND	Internet users (per 100 people)	IT.NET.USER.P2	86.434425	94.000000	95.900000	96.910000	NaN	NaN	IT.NET.USER.P2	Infrastructure Communication

ponderation :

- 1 Gov exp secondary \* 2
- 2 Gov exp tertiary \* 2
- 3 Internet user/100 \* 3
- 4 Growth Rate \* 1
- 5 Pop 15-24 \* 5
- 6 Pop 15-60 \* 2
- 7 % Pop 25+ lowsec \* 3

Les pays dans lesquels investir !







## **Conclusion**

### **Conclusion**

**Modèle pondérable**

**Sélection d'une quinzaine de pays**

### **Limitation de l'analyse**

**Limitation Technique**

**Le modèle**

**La pondération**

**Limitation métier**

**Connaissance du domaine**

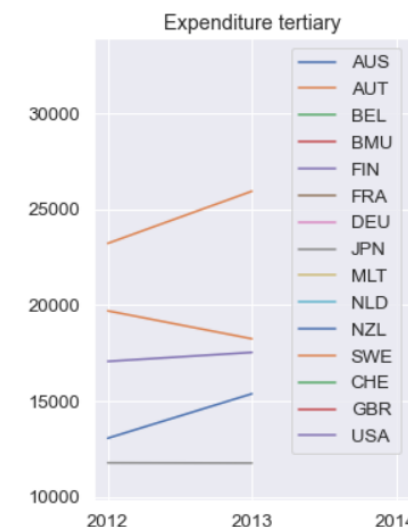
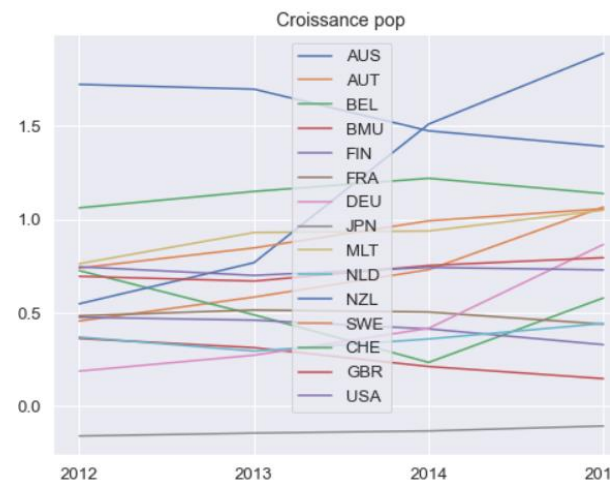
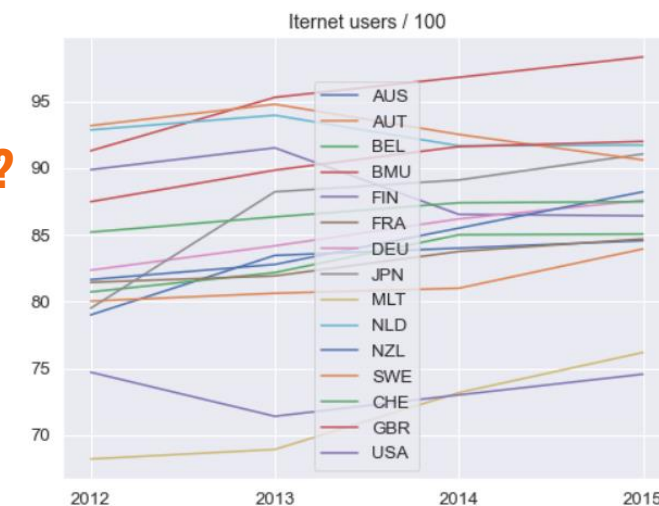
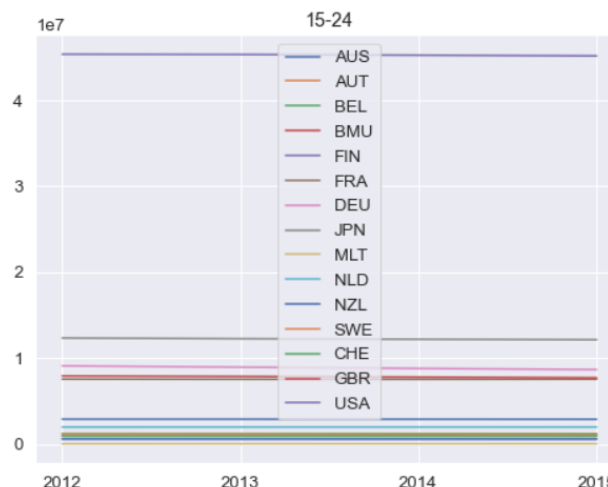
**Sélection des indicateurs**



## La réponse aux 3 questions

- 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 ?

Australia, **Austria**,  
Belgium, Bermuda,  
Finland, France,  
Germany,  
Japan,  
Malta,  
Netherlands, New Zealand,  
**Sweden, Switzerland**,  
**United Kingdom, United States**





**Fin de la présentation**

**Merci de m'avoir écouté !**