Soutenance Projet P2

Objectifs

- Réaliser une analyse pré-exploratoire d'un jeu de données
- Déterminer les colonnes et variables nécessaire à notre analyse
- Calculer différent indicateur statistique pour chaque pays
- BUT : déterminer quels pays possèdent potentiellement le plus de client pour notre entreprise

Base de données

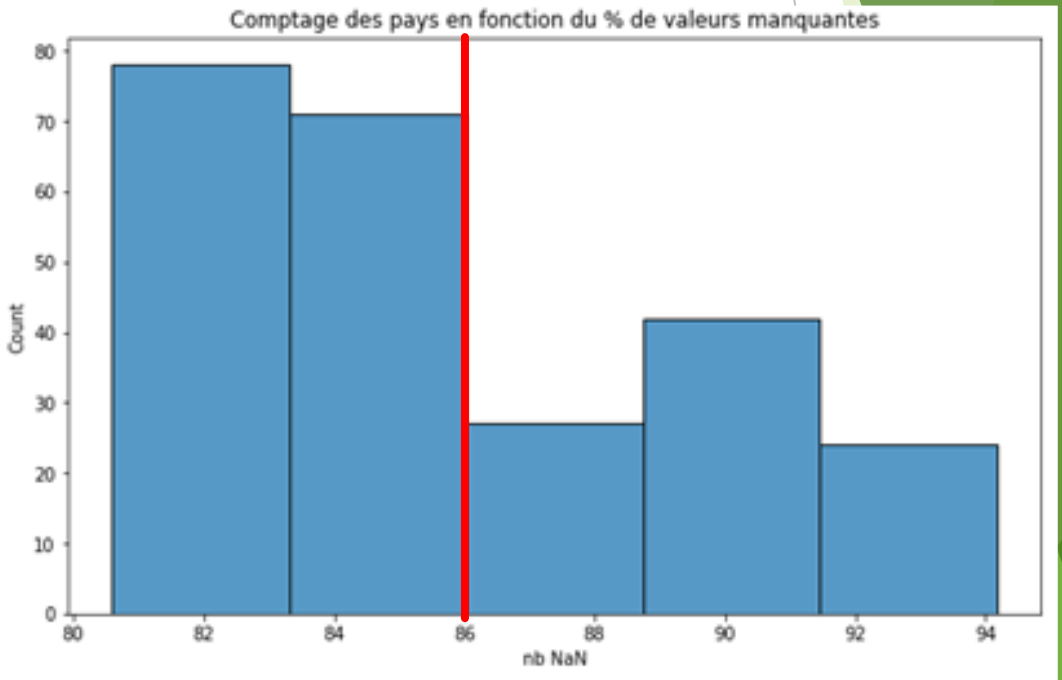
- Taille : 886930 x 69
- 242 pays
- 3665 variables
- Pas de variables dupliqués
- 86% de valeurs manquantes

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...
...	...	...	...	...	...	...	...	...	...	...	...

Nettoyage des données

1er tri des Pays :

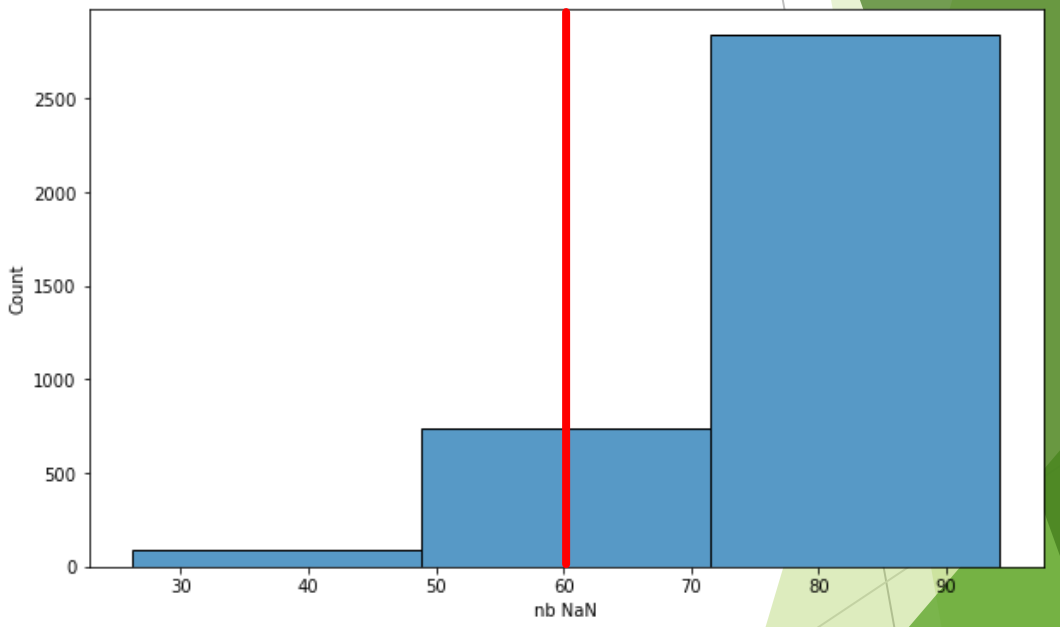
- On détermine le nombre de valeurs manquantes pour chaque pays
- On supprimes les pays dépassant un seuil de 86% de valeurs manquantes



Nettoyage des données

Choix des variables:

- On détermine le nombre de valeurs manquantes pour chaque variables
- On supprimes les variables dépassant un seuil de 60% de valeurs manquantes



Nettoyage des données

Choix des variables:

- Le % des 15-64 ans
- Population en âge officiel pour l'enseignement secondaire supérieur
- Population en âge officiel pour l'enseignement tertiaire
- Utilisateur d'Internet (pour 100 people)

	nom Var	nb NaN
UIS.THAGE.0	Official entrance age to pre-primary education...	26.194673
SP.POP.TOTL.FE.ZS	Population, female (% of total)	26.282805
SP.POP.1564.TO.ZS	Population, ages 15-64 (% of total)	26.282805
SP.POP.0014.TO.ZS	Population, ages 0-14 (% of total)	26.282805
SP.POP.TOTL.MA.ZS	Population, male (% of total)	26.282805
...	...	...
UIS.E.1.PR	Enrolment in primary education, private instit...	59.488837
SE.PRE.TCHR	Teachers in pre-primary education, both sexes ...	59.576968
UIS.NE.1.G1.F	New entrants to Grade 1 of primary education, ...	59.625930
UIS.OAEP.1	Percentage of students enrolled in primary edu...	59.655307
SE.PRE.ENRL.TC.ZS	Pupil-teacher ratio in pre-primary education (...	59.949080

355 rows × 2 columns

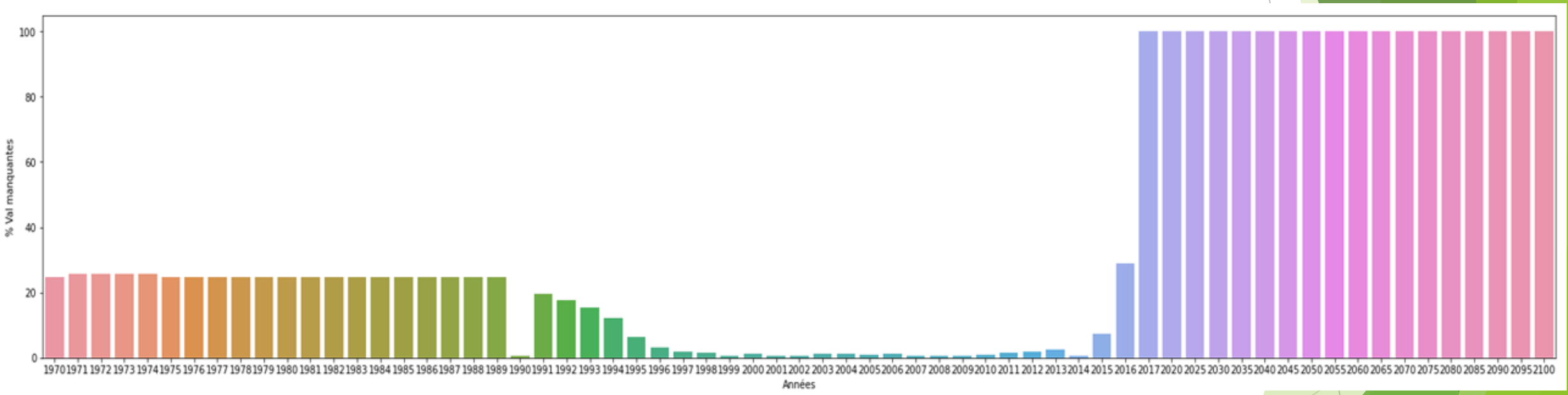
Nettoyage des données

Choix des colonnes:

- On souhaite garder les colonnes les plus pertinentes
- Pour chaque colonnes on détermine le % de valeurs manquantes
- On supprimes les colonnes avec le plus de valeurs manquantes

Nettoyage des données

Choix des colonnes:



Nettoyage des données

2ème tri des Pays :

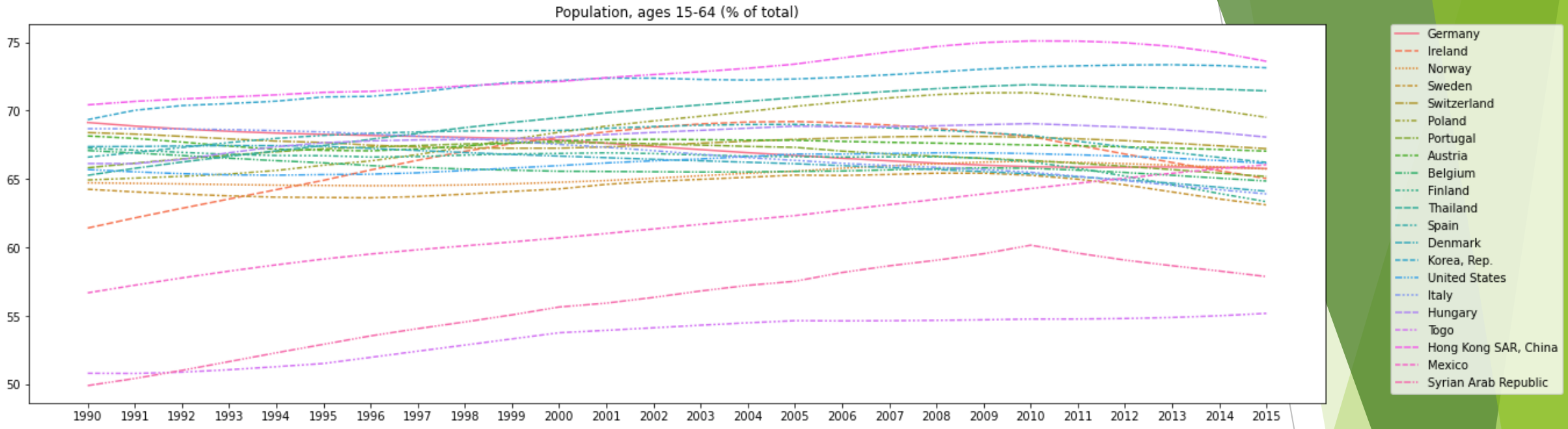
- On détermine le nombre de valeurs manquantes pour chaque pays
- On va garder les pays pour lesquelles on a pas de valeurs manquantes

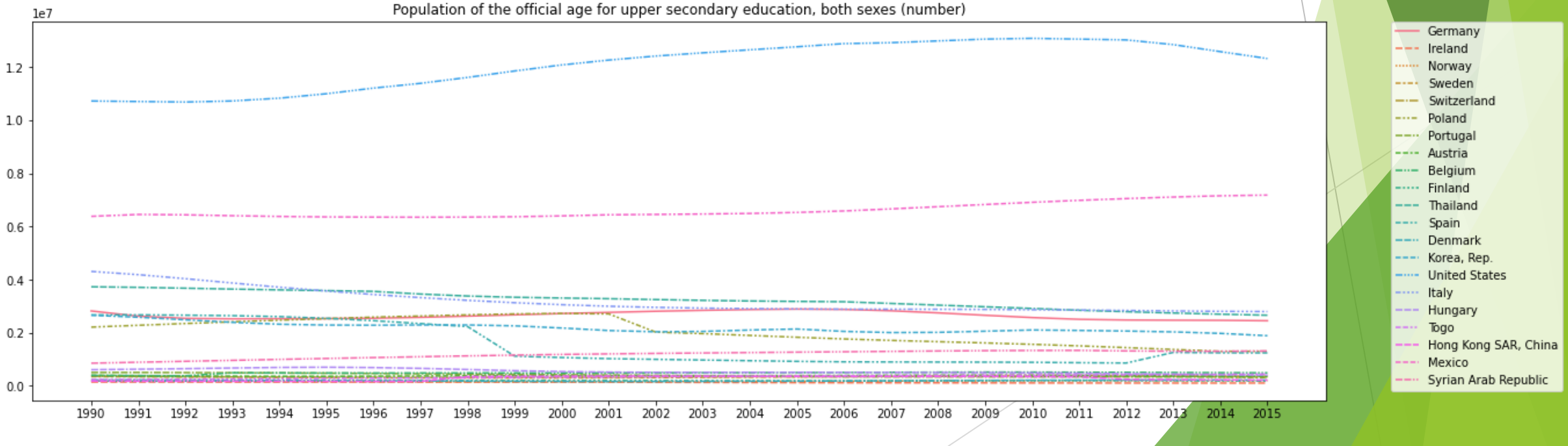
	Pays	% Val manquantes
90	Germany	0.0
28	Ireland	0.0
25	Norway	0.0
23	Sweden	0.0
22	Switzerland	0.0
51	Poland	0.0
53	Portugal	0.0
54	Austria	0.0
32	Belgium	0.0
15	Finland	0.0
107	Thailand	0.0
12	Spain	0.0
35	Denmark	0.0
10	Korea, Rep.	0.0
71	United States	0.0
2	Italy	0.0
4	Hungary	0.0
103	Togo	0.0
66	Hong Kong SAR, China	0.0
7	Mexico	0.0
9	Syrian Arab Republic	0.0

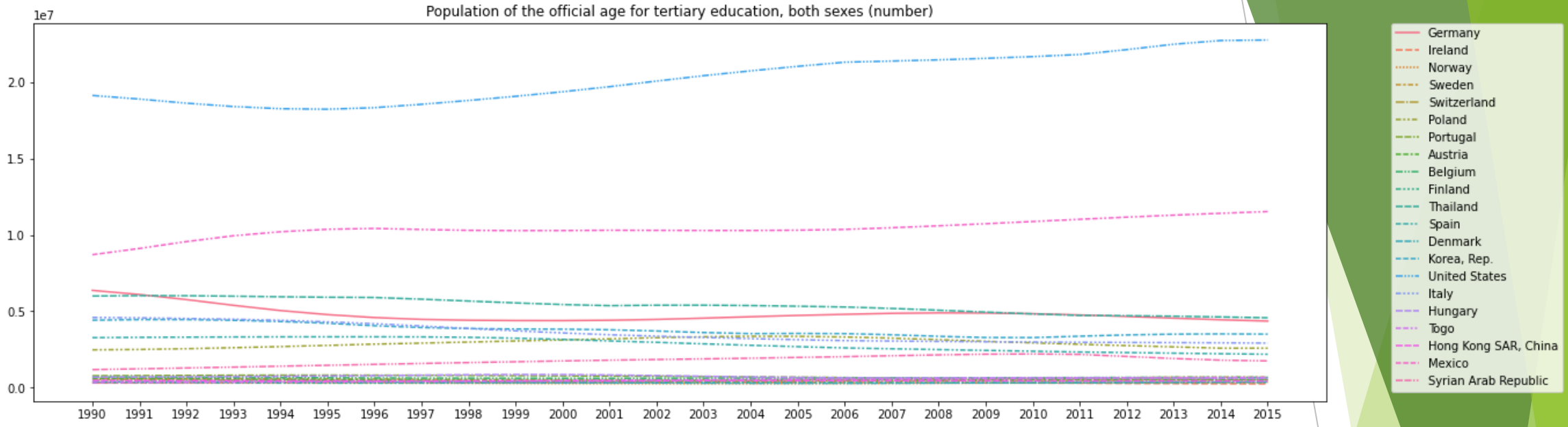
Analyse des données

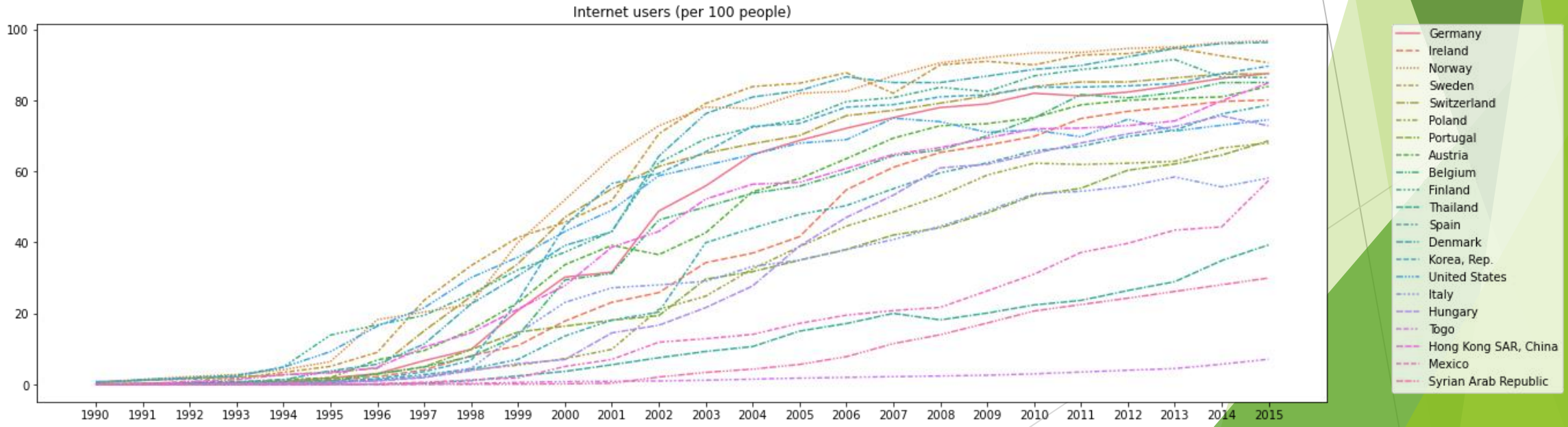
Calculs d'indicateurs statistique

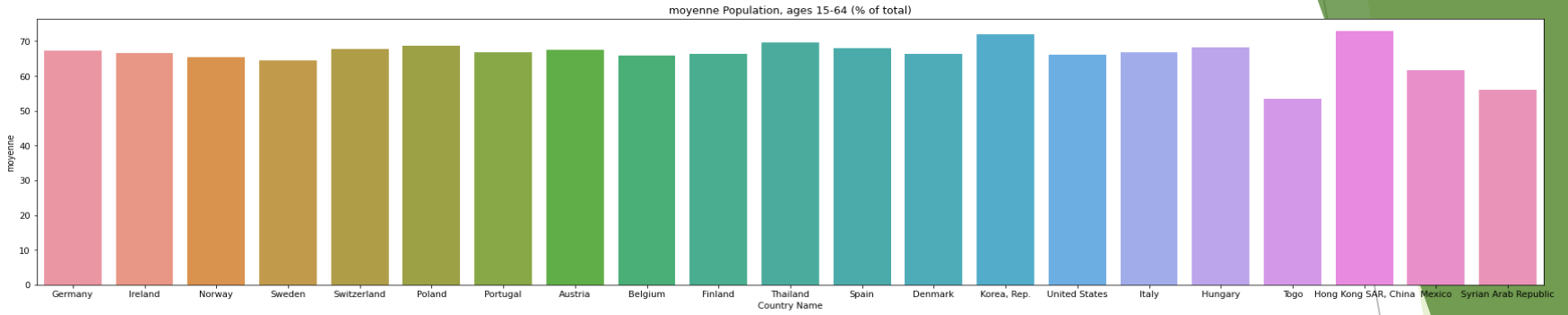
- On calcule la moyenne, variance, etc … pour chaque variables
- On réalise un graphique montrant l'évolution pour chaque variable
- On réalise un diagramme des moyennes de chaque variable

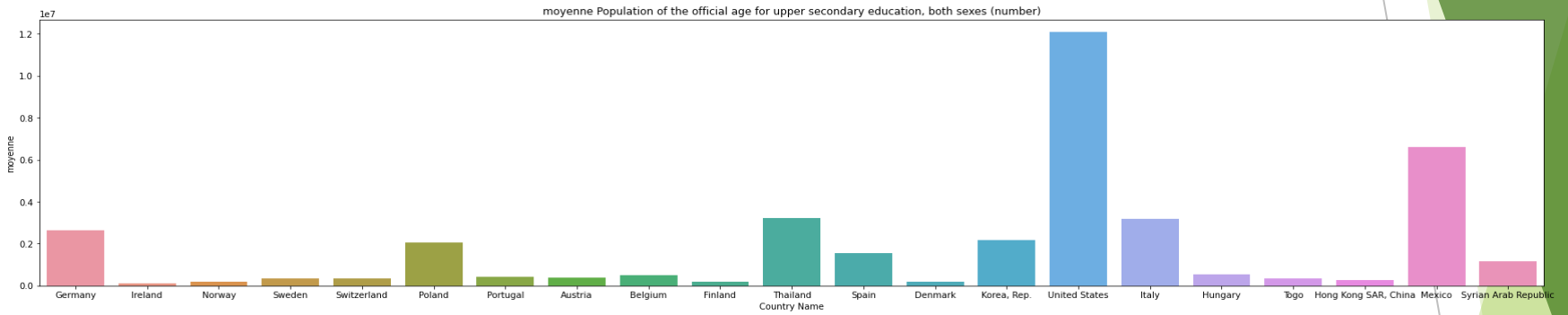

Population, ages 15-64 (% of total)
75
70
65
60
55
50
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic

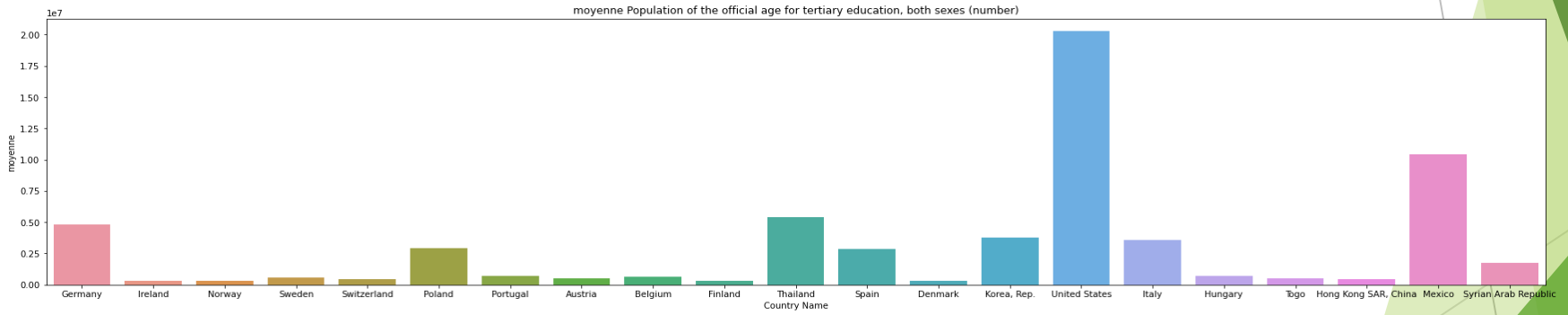

Population of the official age for upper secondary education, both sexes (number)
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic

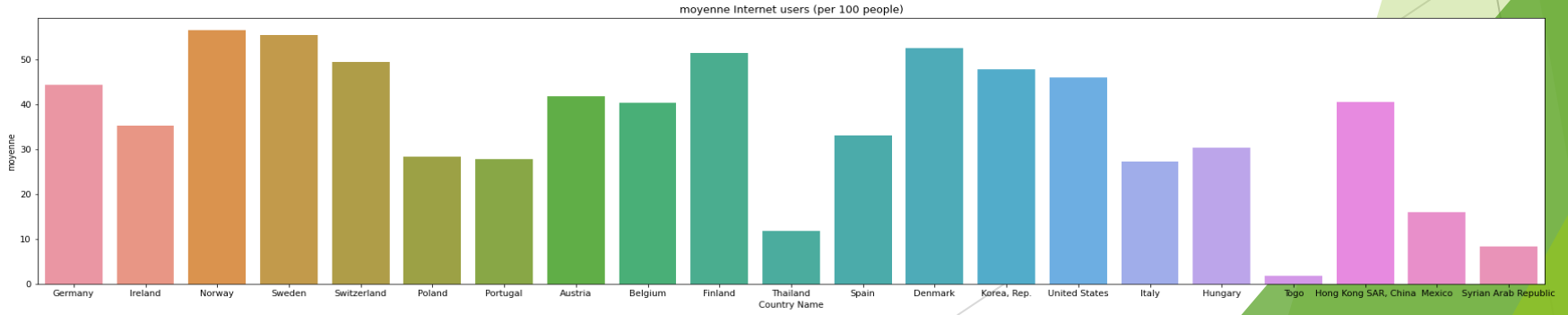

Population of the official age for tertiary education, both sexes (number)
1e7
2.0
1.5
1.0
0.5
0.0
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic


Internet users (per 100 people)
100
80
60
40
20
0
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
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic


moyenne Population, ages 15-64 (% of total)
moyenne
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic
Country Name


moyenne Population of the official age for upper secondary education, both sexes (number)
1e7
moyenne
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic
Country Name


moyenne Population of the official age for tertiary education, both sexes (number)
1e7
moyenne
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic
Country Name


moyenne Internet users (per 100 people)
moyenne
Germany
Ireland
Norway
Sweden
Switzerland
Poland
Portugal
Austria
Belgium
Finland
Thailand
Spain
Denmark
Korea, Rep.
United States
Italy
Hungary
Togo
Hong Kong SAR, China
Mexico
Syrian Arab Republic
Country Name

Conclusion

Les Pays les plus intéressants :

- Mexique
- Allemagne
- Pologne
- Italie

- Jeu de données avec beaucoup de valeurs manquantes
- Nettoyage stricte pour récupérer les pays sans valeurs manquantes