PRESENTATION P2 Data Scientist

Oumeima EL GHARBI, novembre 2021

PLAN

Problématique :

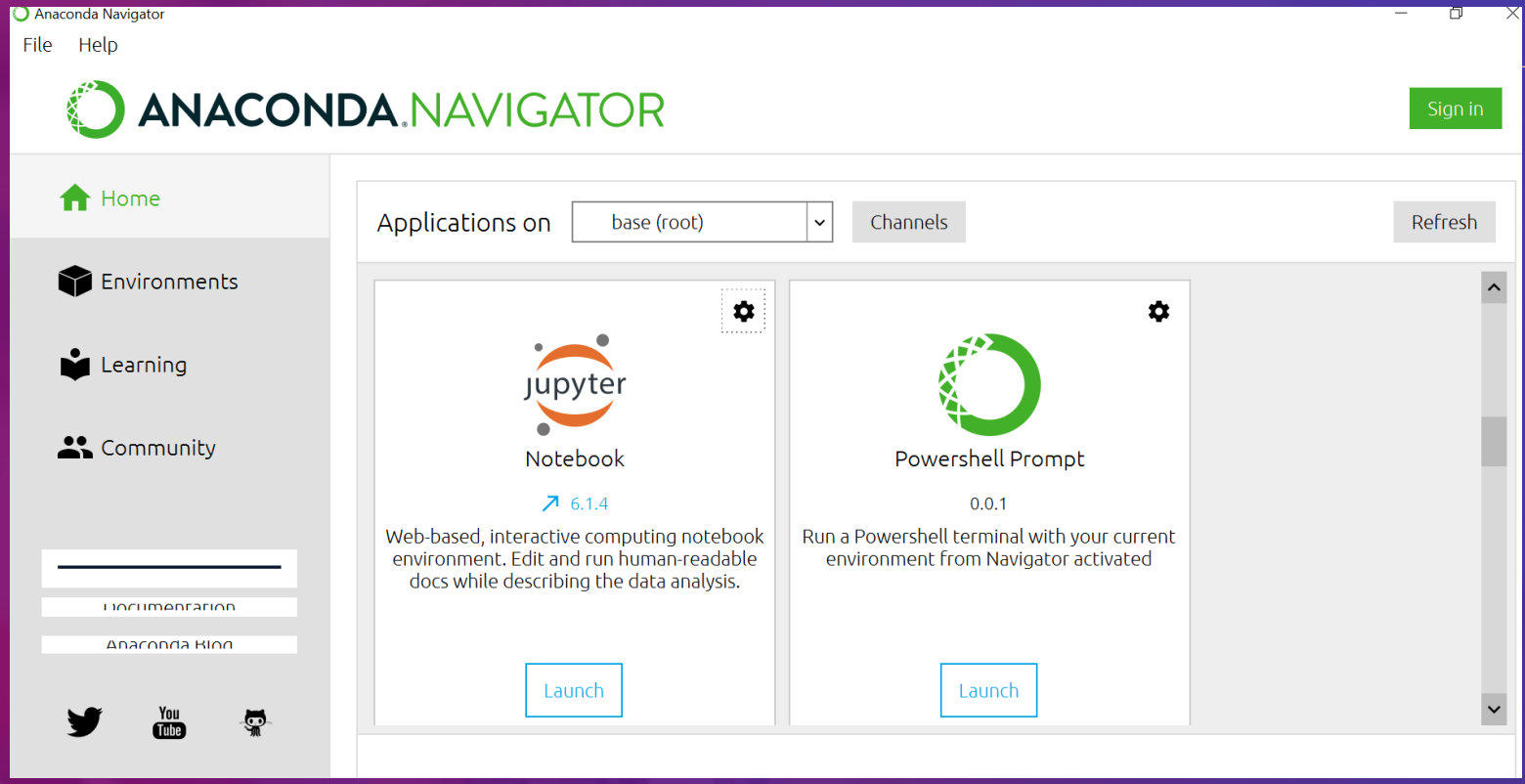
Quels sont les pays avec un fort potentiel de clients pour nos services ?
Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
Dans quels pays l'entreprise doit-elle opérer en priorité ?

Question : Quelles sont les variables permettant de quantifier le potentiel d'un pays en vue d'un développement commercial ?

Objectif : proposer une liste de pays correspondants aux critères de l'entreprise.

INTRODUCTION

Utilisation de deux notebooks Jupyter.



INTRODUCTION

I - Première exploration des données brutes

Premier notebook : P2_01_nettoyage

Cinq fichiers au format csv

Fichiers csv renommés.

Premier contrôle visuel de ces cinq dataframes avec la librairie Pandas.

```
df1 = pd.read_csv('csv/EdStatsCountry.csv')
df2 = pd.read_csv('csv/EdStatsCountry_Series.csv')
df3 = pd.read_csv('csv/EdStatsData.csv')
df4 = pd.read_csv('csv/EdStatsFootNote.csv')
df5 = pd.read_csv('csv/EdStatsSeries.csv')
```

I – Première exploration des données brutes

Df1

Nous affichons le dataframe dans le notebook et constatons que les colonnes ne sont pas toutes pertinentes pour notre problématique.

Nous gardons ainsi les colonnes intitulées “Country Code”, “Short Name” et “Income group” qui nous permettent d’identifier le pays et son niveau de revenu.

Sur les 241 lignes du dataframes, il y en 27 pour lesquelles nous n’avons pas de données pour “Income Group” et nous avons 75 pays à “hauts revenus”.

Remarque : avant de retirer les 27 lignes pour lesquelles nous n’avions pas de valeurs, un contrôle au cas par cas a été fait pour contrôler que ces lignes n’étaient pas pertinentes pour notre étude

I - Première exploration des données brutes

Df1

Nous avons donc effectué un premier filtrage sur le premier dataframe.

Nous conservons ce dataframe qui sera enregistré sous le nom “df_richesse”, il contient 75 pays.

	Country Code	Short Name	Income Group
0	ABW	Aruba	High income: nonOECD
4	AND	Andorra	High income: nonOECD
6	ARE	United Arab Emirates	High income: nonOECD
10	ATG	Antigua and Barbuda	High income: nonOECD
11	AUS	Australia	High income: OECD
...	...	...	...
210	TCA	Turks and Caicos Islands	High income: nonOECD
218	TTO	Trinidad and Tobago	High income: nonOECD
226	URY	Uruguay	High income: nonOECD
227	USA	United States	High income: OECD
231	VIR	Virgin Islands	High income: nonOECD

75 rows × 3 columns

I - Première exploration des données brutes

Df2

Après un contrôle visuel des colonnes de ce dataframe, nous constatons que les données ne sont pas pertinentes pour notre étude, nous décidons donc de ne pas conserver ce dataframe.

	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population Prospects	NaN
1	ABW	SP.POP.GROW	Data sources: United Nations World Population Prospects	NaN
2	AFG	SP.POP.GROW	Data sources: United Nations World Population Prospects	NaN
3	AFG	NY.GDP.PCAP.PP.CD	Estimates are based on regression.	NaN
4	AFG	SP.POP.TOTL	Data sources : United Nations World Population Prospects	NaN

I - Première exploration des données brutes

Df3

Nous affichons le dataframe dans le notebook et constatons que ce dataframe est intéressant car il contient tous les indicateurs (4000 environs).

Les colonnes ne sont pas toutes pertinentes pour notre problématique. En effet, nous ne garderons pas les colonnes "années" avant 2010 (absence d'internet) et après 2035 (projections trop éloignées dans le temps).

Nous gardons ainsi les colonnes intitulées "Country Code", "Country Name", "Indicator Name", et les colonnes de "2010" à "2030".

	Country Name	Country Code	Indicator Name	2010	2011	2012	2013	2014	2015	2016	2020	2025	2030
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, both sexes (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, female (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, gender parity index (GPI)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, male (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sexes (%)	85.211998	85.24514	86.101669	85.51194	85.320152	NaN	NaN	NaN	NaN	NaN

I - Première exploration des données brutes

Df3

Deuxième étape pour le df3 : nous ne gardons que les 75 pays à hauts revenus retenus dans le df_richesse.

Enregistement de ce dataframe : df_indicators.csv

	Country Name	Country Code	Indicator Name	2010	2011	2012	2013	2014	2015	2016	2020	2025	2030
106285	Andorra	AND	Adjusted net enrolment rate, lower secondary, both sexes (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
106286	Andorra	AND	Adjusted net enrolment rate, lower secondary, female (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
106287	Andorra	AND	Adjusted net enrolment rate, lower secondary, gender parity index (GPI)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
106288	Andorra	AND	Adjusted net enrolment rate, lower secondary, male (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
106289	Andorra	AND	Adjusted net enrolment rate, primary, both sexes (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
872265	Virgin Islands (U.S.)	VIR	Youth illiterate population, 15-24 years, male (number)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
872266	Virgin Islands (U.S.)	VIR	Youth literacy rate, population 15-24 years, both sexes (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
872267	Virgin Islands (U.S.)	VIR	Youth literacy rate, population 15-24 years, female (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
872268	Virgin Islands (U.S.)	VIR	Youth literacy rate, population 15-24 years, gender parity index (GPI)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
872269	Virgin Islands (U.S.)	VIR	Youth literacy rate, population 15-24 years, male (%)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

I - Première exploration des données brutes

Df4

Après un contrôle visuel des colonnes de ce dataframe, nous constatons que les données ne sont pas pertinentes pour notre étude, nous décidons donc de ne pas conserver ce dataframe.

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN

I – Première exploration des données brutes

Df5

Après un contrôle visuel des colonnes de ce dataframe, nous constatons que les données ne sont pas pertinentes pour notre étude, nous décidons donc de ne pas conserver ce dataframe.

Cependant, ce dataframe contient les définitions des indicateurs du df3, nous avons donc essayé d'utiliser ce dataframe pour obtenir plus de precision sur les indicateurs : pas pertinent finalement.

II - Choix des indicateurs

1) Contrôle visuel

Nous choisissons 30 indicateurs pertinents pour notre analyse exploratoire.

Cela se fera avec un contrôle visuel des 4000 indicateurs du jeu de données 3.

II - Choix des indicateurs

2) Méthode de nettoyage

Nettoyer df_indicators en fonction de :

A) nombre d'habitants par pays.

L'indicateur :"Population, ages 15-64, total"

B) Taux d'utilisation d'internet pour 100 habitants

L'ndicateur : "'Internet users (per 100 people)"

C) Niveau d'éducation de la population : 11 indicateurs retenus.

II - Choix des indicateurs

2) A) Population

Nous constatons que sur les cinq indicateurs de population, 'Population, ages 15-64, total' a le meilleur taux de remplissage : 60/75 % soit de 80%.

Nous gardons la colonne 2016 étant la plus proche de 2020.

Nous ne gardons donc que cet indicateur et nous allons retirer les pays trop petits pour l'implémentation d'un établissement de formation en ligne. Nous prenons arbitrairement 10 millions d'habitants comme seuil minimum.

Nous utiliserons un autre indicateur pour le nombre de jeunes entre 15 et 30 ans dans ces pays.

II - Choix des indicateurs

2) A) Population

Remarque : Nous avons vérifié que les NaN pour 2016 correspondaient aux petits pays sur les 75 pays riches pré-sélectionnés.

	Country Name	Country Code	Indicator Name	2010	2011	2012	2013	2014	2015	2016	2020	2025	2030
2486	Andorra	AND	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
35471	Bermuda	BMU	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
46466	Cayman Islands	CYM	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
61126	Curacao	CUW	Population, ages 15-64, total	100007.0	100766.0	101579.0	102283.0	NaN	NaN	NaN	NaN	NaN	NaN
83116	Faroe Islands	FRO	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
105106	Greenland	GRL	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
123431	Isle of Man	IMN	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
149086	Liechtenstein	LIE	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
167411	Monaco	MCO	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
182071	Northern Mariana Islands	MNP	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
211391	San Marino	SMR	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
222386	Sint Maarten (Dutch part)	SXM	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
237046	St. Kitts and Nevis	KNA	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
240711	St. Martin (French part)	MAF	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
255371	Turks and Caicos Islands	TCA	Population, ages 15-64, total	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

II - Choix des indicateurs

2) A) Population

Il nous reste donc 15 pays riches de plus de 10 millions d'habitants : df_population.csv

Nous attribuons à chaque pays une note allant de 1 à 15 en fonction de leur classement : ils sont classés par ordre décroissant, le plus grand pays ayant 15 et le plus petit ayant 1.

	Country Name	Country Code	Indicator Name	2014	2015	2016
266366	United States	USA	Population, ages 15-64, total	211378325.0	212262832.0	213071223.0
207726	Russian Federation	RUS	Population, ages 15-64, total	101035287.0	100404879.0	99477057.0
134426	Japan	JPN	Population, ages 15-64, total	78379970.0	77547638.0	76831284.0
97776	Germany	DEU	Population, ages 15-64, total	53327698.0	53720119.0	54263836.0
262701	United Kingdom	GBR	Population, ages 15-64, total	41744392.0	41873827.0	42028042.0
90446	France	FRA	Population, ages 15-64, total	41917948.0	41837530.0	41796373.0
130761	Italy	ITA	Population, ages 15-64, total	39035661.0	38813169.0	38591970.0
138091	Korea, Rep.	KOR	Population, ages 15-64, total	37193108.0	37307195.0	37364822.0
233381	Spain	ESP	Population, ages 15-64, total	30946615.0	30755139.0	30678609.0
193066	Poland	POL	Population, ages 15-64, total	26612967.0	26402284.0	26187409.0
42801	Canada	CAN	Population, ages 15-64, total	24258899.0	24332020.0	24477514.0
215056	Saudi Arabia	SAU	Population, ages 15-64, total	21682073.0	22389217.0	23013048.0
13481	Australia	AUS	Population, ages 15-64, total	15598904.0	15745617.0	15887445.0
53796	Chile	CHL	Population, ages 15-64, total	12107339.0	12210505.0	12304561.0
171076	Netherlands	NLD	Population, ages 15-64, total	11068561.0	11065869.0	11068819.0

II - Choix des indicateurs

2) A) Population

Classement :

{'United States': 15, 'Russian Federation': 14, 'Japan': 13,

'Germany': 12, 'United Kingdom': 11, 'France': 10,

'Italy': 9, 'Korea, Rep.': 8, 'Spain': 7,

'Poland': 6, 'Canada': 5, 'Saudi Arabia': 4,

'Australia': 3, 'Chile': 2, 'Netherlands': 1}

II - Choix des indicateurs

2) B) Internet

Parmi les 25 indicateurs restants, on s'intéresse au taux d'utilisateurs d'internet pour 100 habitants pour chacun des 15 pays pré-sélectionnés.

Le taux de remplissage est de 100% pour ces 15 pays riches et de plus de 10 millions d'habitants pour les années de 2010 à 2016. Nous trions par ordre décroissant pour l'année 2016 puis pour 2015.

Nous avons un nouveau dataframe : df_internet.csv

II - Choix des indicateurs

2) B) Internet

Nous attribuons à nouveau une note allant de 1 à 15 pour chaque pays qui est sommée avec la première note :

{'United States': 20, 'Russian Federation': 20, 'Japan': 26, 'Germany': 22, 'United Kingdom': 26, 'France': 18, 'Italy': 10, 'Korea, Rep.': 22, 'Spain': 14, 'Poland': 9, 'Canada': 16, 'Saudi Arabia': 8, 'Australia': 12, 'Chile': 4, 'Netherlands': 13}

	Country Name	Country Code	Indicator Name	2015	2016
261590	United Kingdom	GBR	Internet users (per 100 people)	92.000300	94.775801
136980	Korea, Rep.	KOR	Internet users (per 100 people)	89.648631	92.716545
133315	Japan	JPN	Internet users (per 100 people)	91.058028	92.000000
169965	Netherlands	NLD	Internet users (per 100 people)	91.724138	90.410959
41690	Canada	CAN	Internet users (per 100 people)	88.470000	89.840000
96665	Germany	DEU	Internet users (per 100 people)	87.589800	89.647101
12370	Australia	AUS	Internet users (per 100 people)	84.560519	88.238658
89335	France	FRA	Internet users (per 100 people)	84.694500	85.622200
232270	Spain	ESP	Internet users (per 100 people)	78.689600	80.561333
206615	Russian Federation	RUS	Internet users (per 100 people)	73.410000	76.409085
265255	United States	USA	Internet users (per 100 people)	74.554202	76.176737
213945	Saudi Arabia	SAU	Internet users (per 100 people)	69.616236	73.750904
191955	Poland	POL	Internet users (per 100 people)	67.997000	73.300700
52685	Chile	CHL	Internet users (per 100 people)	64.289000	66.010000
129650	Italy	ITA	Internet users (per 100 people)	58.141735	61.324253

II - Choix des indicateurs

2) C) Education

- Il nous reste 24 indicateurs sur les 30 indicateurs initialement choisis.

- Nous étudions avec DataFrame.info() le taux de remplissage pour chaque indicateurs.

- Nous ne conservons que 11 indicateurs ayant des taux de remplissage supérieur à 50% par colonne "année".

- Pour chacun de ses indicateurs : nous avons classés les pays par ordre décroissant.

- Dans le cas où il y a une valeur manquante pour un pays, nous avons choisi de remplir la donnée absente par la donnée de l'année precedent ou via des recherches (Canada).

II - Choix des indicateurs

2) C) Education

- Imputation par la méthode par le plus proche voisin : on attribue à l'enregistrement pour lequel la réponse à une question manque la valeur figurant pour cette question dans l'enregistrement obtenu pour le répondant le plus proche.

- En voici un exemple avec le premier des 11 indicateurs.

	Country Name	Country Code	Indicator Name	2013	2014
11003	Australia	AUS	Adjusted net enrolment rate, upper secondary, both sexes (%)	76.045959	NaN
40323	Canada	CAN	Adjusted net enrolment rate, upper secondary, both sexes (%)	84.529091	NaN
51318	Chile	CHL	Adjusted net enrolment rate, upper secondary, both sexes (%)	82.885231	82.265427
87968	France	FRA	Adjusted net enrolment rate, upper secondary, both sexes (%)	87.904182	88.872711
95298	Germany	DEU	Adjusted net enrolment rate, upper secondary, both sexes (%)	NaN	NaN
128283	Italy	ITA	Adjusted net enrolment rate, upper secondary, both sexes (%)	88.781151	NaN
131948	Japan	JPN	Adjusted net enrolment rate, upper secondary, both sexes (%)	97.071823	NaN
135613	Korea, Rep.	KOR	Adjusted net enrolment rate, upper secondary, both sexes (%)	91.108917	NaN
168598	Netherlands	NLD	Adjusted net enrolment rate, upper secondary, both sexes (%)	56.670990	56.299019
190588	Poland	POL	Adjusted net enrolment rate, upper secondary, both sexes (%)	85.990601	NaN
205248	Russian Federation	RUS	Adjusted net enrolment rate, upper secondary, both sexes (%)	NaN	NaN
212578	Saudi Arabia	SAU	Adjusted net enrolment rate, upper secondary, both sexes (%)	61.348930	69.297653
230903	Spain	ESP	Adjusted net enrolment rate, upper secondary, both sexes (%)	78.541397	80.409363
260223	United Kingdom	GBR	Adjusted net enrolment rate, upper secondary, both sexes (%)	88.727089	92.257690
263888	United States	USA	Adjusted net enrolment rate, upper secondary, both sexes (%)	77.483688	79.888199

II - Choix des indicateurs

2) C) Education

- Voici un exemple avec le premier indicateur :

Nous remplissons pour 2014 avec la valeur de 2013 pour cet indicateur « Adjusted net enrolment rate, upper secondary, both sexes (%) ».

- Nous avons procédé de la même marnière pour les 10 autres indicateurs.

- Nous nous basons pour le classement sur l'année la plus proche de 2021.

	Country Name	Country Code	Indicator Name	2013	2014
131948	Japan	JPN	Adjusted net enrolment rate, upper secondary, both sexes (%)	97.071823	97.071823
260223	United Kingdom	GBR	Adjusted net enrolment rate, upper secondary, both sexes (%)	88.727089	92.257690
135613	Korea, Rep.	KOR	Adjusted net enrolment rate, upper secondary, both sexes (%)	91.108917	91.108917
87968	France	FRA	Adjusted net enrolment rate, upper secondary, both sexes (%)	87.904182	88.872711
128283	Italy	ITA	Adjusted net enrolment rate, upper secondary, both sexes (%)	88.781151	88.781151
190588	Poland	POL	Adjusted net enrolment rate, upper secondary, both sexes (%)	85.990601	85.990601
40323	Canada	CAN	Adjusted net enrolment rate, upper secondary, both sexes (%)	84.529091	84.529091
51318	Chile	CHL	Adjusted net enrolment rate, upper secondary, both sexes (%)	82.885231	82.265427
230903	Spain	ESP	Adjusted net enrolment rate, upper secondary, both sexes (%)	78.541397	80.409363
263888	United States	USA	Adjusted net enrolment rate, upper secondary, both sexes (%)	77.483688	79.888199
11003	Australia	AUS	Adjusted net enrolment rate, upper secondary, both sexes (%)	76.045959	76.045959
212578	Saudi Arabia	SAU	Adjusted net enrolment rate, upper secondary, both sexes (%)	61.348930	69.297653
168598	Netherlands	NLD	Adjusted net enrolment rate, upper secondary, both sexes (%)	56.670990	56.299019
95298	Germany	DEU	Adjusted net enrolment rate, upper secondary, both sexes (%)	0.000000	0.000000
205248	Russian Federation	RUS	Adjusted net enrolment rate, upper secondary, both sexes (%)	0.000000	0.000000

II - Choix des indicateurs

2) C) Education : choix de 4 indicateurs :

- Enrolment in upper secondary education, both sexes (number)
- Enrolment in tertiary education, all programmes, both sexes (number) : biaisé car supérieur pour les pays ayant une grande population, manque une donnée : Canada.
- Gross enrolment ratio, tertiary, both sexes (%) : nous donne le pourcentage d'étudiants poursuivant dans l'enseignement supérieur, manque une donnée : Canada.
- Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total

Nous choisissons les indicateurs avec des nombres et non pas des pourcentages car nous ciblons les pays avec une grande population d'étudiants (enseignement supérieur, et enseignement secondaire car ce sont les étudiants de demain).

II - Choix des indicateurs

2) C) Education : "nb_inscrits_secondaire.csv"

- Enrolment in upper secondary education, both sexes (number)

Il s'agit du nombre total d'étudiants inscrits dans l'enseignement secondaire : ce sont les étudiants de demain.

Taux de remplissage 100%, données anciennes.

	Country Name	Country Code	Indicator Name	2013	2014
265094	United States	USA	Enrolment in upper secondary education, both sexes (number)	11646415.00	11736315.00
261429	United Kingdom	GBR	Enrolment in upper secondary education, both sexes (number)	4117193.00	4195081.50
133154	Japan	JPN	Enrolment in upper secondary education, both sexes (number)	3682920.00	3682920.00
206454	Russian Federation	RUS	Enrolment in upper secondary education, both sexes (number)	2972383.00	2823004.00
129489	Italy	ITA	Enrolment in upper secondary education, both sexes (number)	2780440.00	2780440.00
89174	France	FRA	Enrolment in upper secondary education, both sexes (number)	2581511.00	2598357.00
96504	Germany	DEU	Enrolment in upper secondary education, both sexes (number)	2575681.25	2579952.25
136819	Korea, Rep.	KOR	Enrolment in upper secondary education, both sexes (number)	1903857.00	1903857.00
213784	Saudi Arabia	SAU	Enrolment in upper secondary education, both sexes (number)	1667593.00	1678613.00
232109	Spain	ESP	Enrolment in upper secondary education, both sexes (number)	1632885.00	1662580.00
191794	Poland	POL	Enrolment in upper secondary education, both sexes (number)	1589524.00	1589524.00
41529	Canada	CAN	Enrolment in upper secondary education, both sexes (number)	1531393.00	1531393.00
12209	Australia	AUS	Enrolment in upper secondary education, both sexes (number)	1079568.00	1104162.00
52524	Chile	CHL	Enrolment in upper secondary education, both sexes (number)	1045167.00	1032041.00
169804	Netherlands	NLD	Enrolment in upper secondary education, both sexes (number)	753872.00	746415.00

II - Choix des indicateurs

2) C) Education : "nb_inscrits_superieur.csv"

- Enrolment in tertiary education, all programmes, both sexes (number)

Il s'agit du nombre total d'étudiants inscrits dans l'enseignement supérieur.

Nous n'avons pas de données pour le Canada, mais nous ne le retirons pas car c'est un pays repute pour son niveau d'education : imputation de la donnée manquante via une autre source de données.

	Country Name	Country Code	Indicator Name	2013	2014	2015
265084	United States	USA	Enrolment in tertiary education, all programmes, both sexes (number)	19972624.00	19700220.00	19531728.00
206444	Russian Federation	RUS	Enrolment in tertiary education, all programmes, both sexes (number)	7528163.00	6995732.00	6592416.00
133144	Japan	JPN	Enrolment in tertiary education, all programmes, both sexes (number)	3862749.00	3862460.00	3862460.00
136809	Korea, Rep.	KOR	Enrolment in tertiary education, all programmes, both sexes (number)	3342264.00	3318307.00	3268099.00
96494	Germany	DEU	Enrolment in tertiary education, all programmes, both sexes (number)	2780012.75	2912203.50	2977781.00
89164	France	FRA	Enrolment in tertiary education, all programmes, both sexes (number)	2338135.00	2388880.00	2388880.00
261419	United Kingdom	GBR	Enrolment in tertiary education, all programmes, both sexes (number)	2386199.00	2352932.75	2352932.75
232099	Spain	ESP	Enrolment in tertiary education, all programmes, both sexes (number)	1969413.00	1982162.00	1963924.00
129479	Italy	ITA	Enrolment in tertiary education, all programmes, both sexes (number)	1872693.00	1854360.00	1826477.00
191784	Poland	POL	Enrolment in tertiary education, all programmes, both sexes (number)	1902718.00	1762666.00	1762666.00
213774	Saudi Arabia	SAU	Enrolment in tertiary education, all programmes, both sexes (number)	1356602.00	1496730.00	1527769.00
12199	Australia	AUS	Enrolment in tertiary education, all programmes, both sexes (number)	1390478.00	1453521.00	1453521.00
52514	Chile	CHL	Enrolment in tertiary education, all programmes, both sexes (number)	1174011.00	1205182.00	1221774.00
169794	Netherlands	NLD	Enrolment in tertiary education, all programmes, both sexes (number)	0.00	0.00	842601.00
41519	Canada	CAN	Enrolment in tertiary education, all programmes, both sexes (number)	0.00	0.00	0.00

II - Choix des indicateurs

2) C) Education : "poursuite_etudes_sup.csv"

- Gross enrolment ratio, tertiary, both sexes (%)

- Il s'agit du pourcentage d'étudiants inscrits dans l'enseignement supérieur après le cycle secondaire.

- Donne une idée de la poursuite d'études ou non après l'enseignement secondaire.

- Nous n'avons pas de données pour le Canada, mais nous ne le retirons pas car c'est un pays réputé pour son niveau d'education.

	Country Name	Country Code	Indicator Name	2014	2015
136944	Korea, Rep.	KOR	Gross enrolment ratio, tertiary, both sexes (%)	94.210213	93.179138
12334	Australia	AUS	Gross enrolment ratio, tertiary, both sexes (%)	90.306503	90.306503
232234	Spain	ESP	Gross enrolment ratio, tertiary, both sexes (%)	89.072121	89.670143
52649	Chile	CHL	Gross enrolment ratio, tertiary, both sexes (%)	86.630699	88.577293
265219	United States	USA	Gross enrolment ratio, tertiary, both sexes (%)	86.663963	85.795776
206579	Russian Federation	RUS	Gross enrolment ratio, tertiary, both sexes (%)	78.653374	80.394081
169929	Netherlands	NLD	Gross enrolment ratio, tertiary, both sexes (%)	0.000000	78.501068
96629	Germany	DEU	Gross enrolment ratio, tertiary, both sexes (%)	65.473801	68.265587
191919	Poland	POL	Gross enrolment ratio, tertiary, both sexes (%)	68.113617	68.113617
89299	France	FRA	Gross enrolment ratio, tertiary, both sexes (%)	64.390472	64.390472
133279	Japan	JPN	Gross enrolment ratio, tertiary, both sexes (%)	63.362591	63.362591
213909	Saudi Arabia	SAU	Gross enrolment ratio, tertiary, both sexes (%)	61.112019	63.066219
129614	Italy	ITA	Gross enrolment ratio, tertiary, both sexes (%)	63.095852	62.496071
261554	United Kingdom	GBR	Gross enrolment ratio, tertiary, both sexes (%)	56.476768	56.476768
41654	Canada	CAN	Gross enrolment ratio, tertiary, both sexes (%)	0.000000	0.000000

II - Choix des indicateurs

2) C) Education : "projection_jeunes_diplomes_secondaire.csv"

- Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total

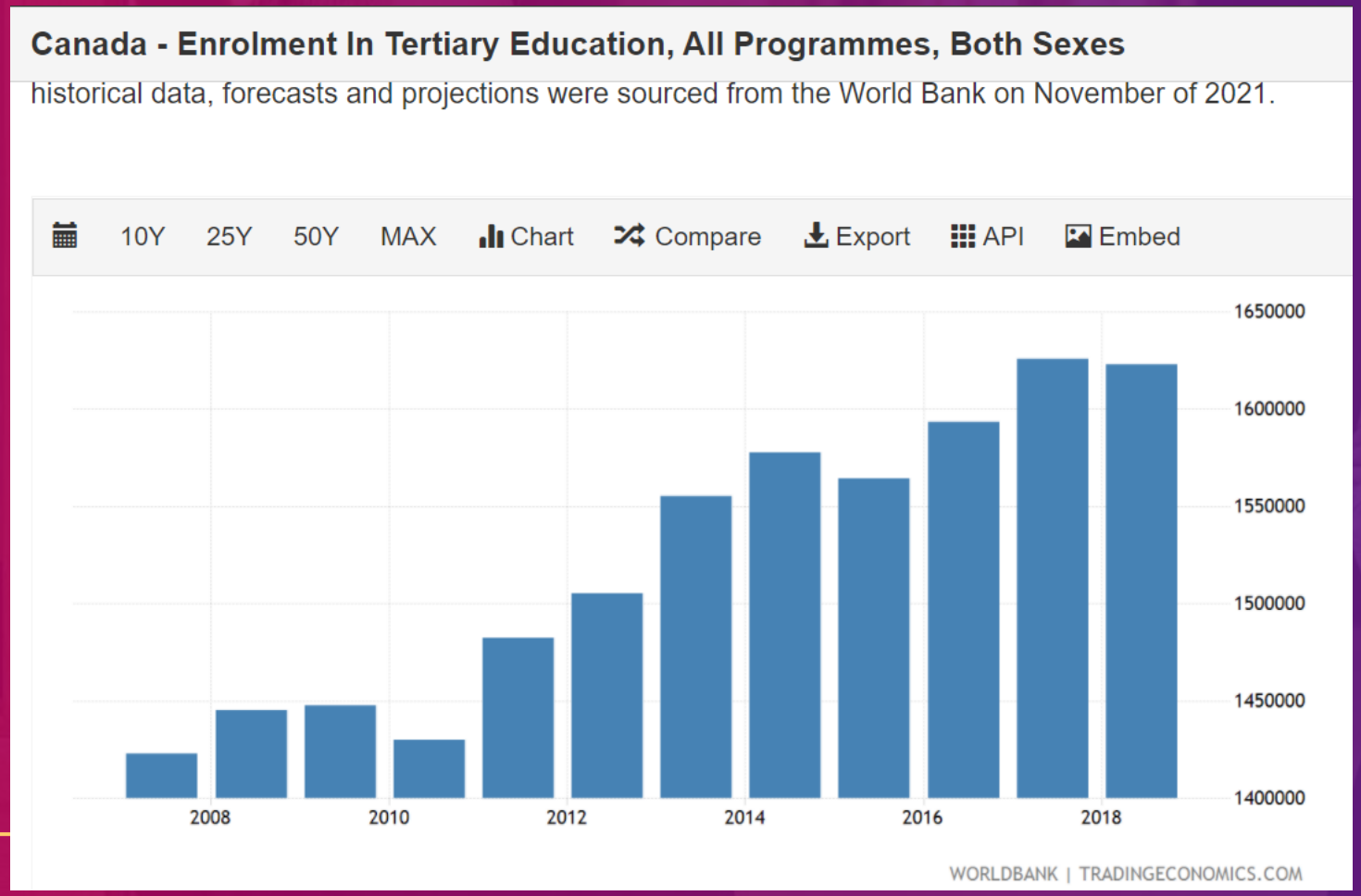
Projection : pourcentage de la population âgée de 20 à 39 ans ayant fini les études secondaires.

Taux de remplissage de 100%, données récentes et futures.

	Country Name	Country Code	Indicator Name	2015	2020	2025	2030
43767	Canada	CAN	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.61	0.64	0.66	0.68
139057	Korea, Rep.	KOR	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.57	0.61	0.66	0.68
135392	Japan	JPN	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.57	0.61	0.65	0.67
91412	France	FRA	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.40	0.43	0.46	0.49
14447	Australia	AUS	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.40	0.43	0.45	0.47
216022	Saudi Arabia	SAU	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.34	0.38	0.42	0.45
263667	United Kingdom	GBR	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.36	0.38	0.41	0.43
98742	Germany	DEU	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.35	0.38	0.40	0.42
267332	United States	USA	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.36	0.38	0.40	0.41
234347	Spain	ESP	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.32	0.35	0.38	0.41
172042	Netherlands	NLD	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.32	0.34	0.37	0.39
194032	Poland	POL	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.26	0.28	0.30	0.31
54762	Chile	CHL	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.21	0.24	0.26	0.28
208692	Russian Federation	RUS	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.22	0.24	0.24	0.25
131727	Italy	ITA	Wittgenstein Projection: Percentage of the population age 20-39 by highest level of educational attainment. Post Secondary. Total	0.16	0.18	0.20	0.22

II – Choix des indicateurs

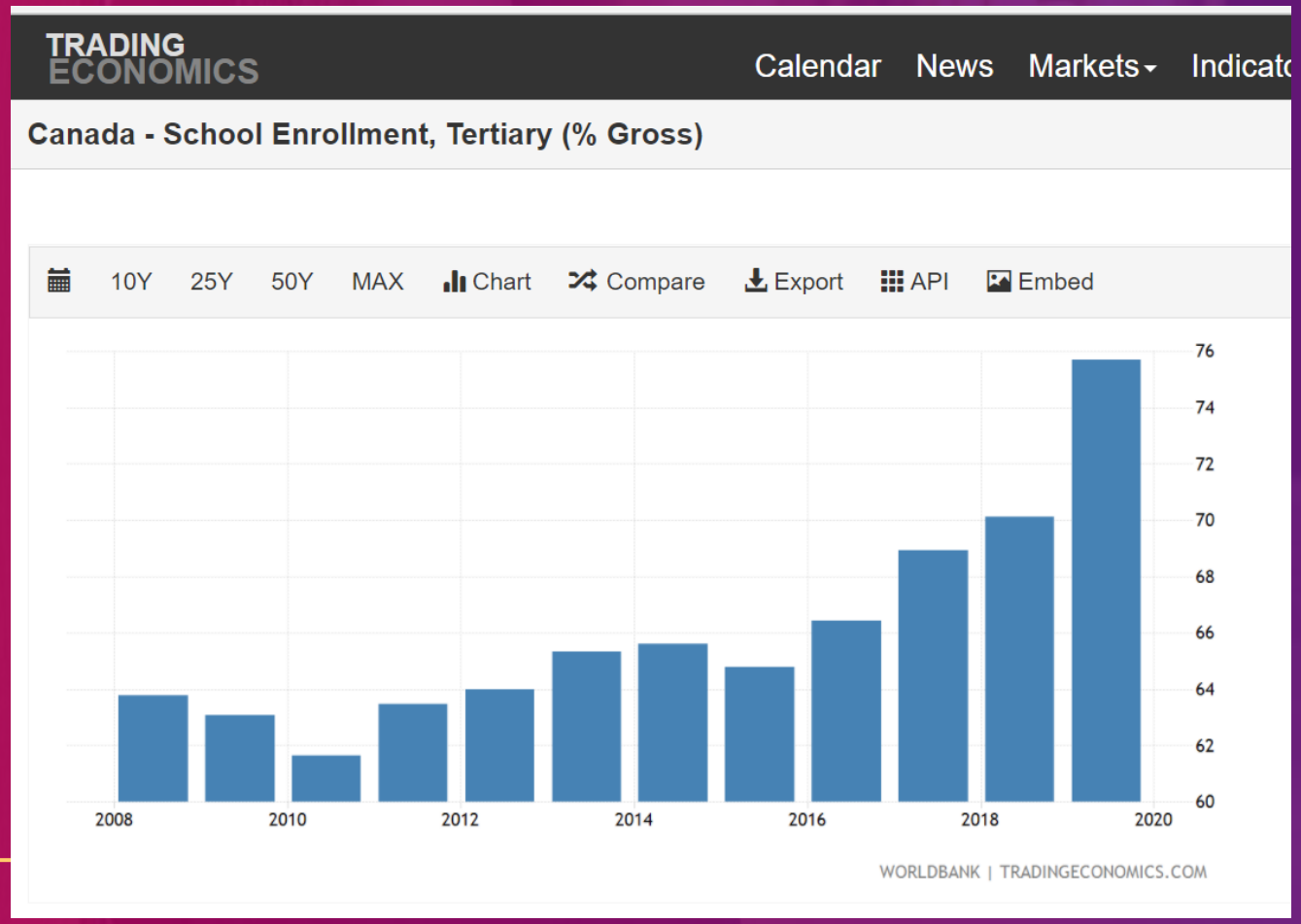
3) Ajustements : 1 -Canada, 2015



Imputation de la valeur de 2015 pour le Canada

II – Choix des indicateurs

3) Ajustements : 2- Canada, 2015



Imputation de la valeur pour 2015 pour le Canada.

II - Choix des indicateurs

4) Score final par pays : 7 pays choisis

Score de la France : 55, gardé ceux supérieur à 55.

Le score est sur 90 car il y a 6 indicateurs et 15 pays.

7 pays choisis :

Corée du Sud : 71

Japon : 69

Etats-Unis : 68

Royaume-Uni : 59

Allemagne : 58

Russie : 58

France : 55

On retire ces 8 pays dont le score est inférieur à celui de la France :

Espagne : 47

Canada : 46

Australie : 43

Arabie Saoudite : 32

Italie : 31

Pologne : 31

Pays-Bas : 29

Chili : 23

III – Visualisation et choix pays

Deuxième notebook : P2_02_exploration

Pays : Corée du Sud, Japon, Etats-Unis, Royaume-Uni, Allemagne, Russie, France.

```
df_population = pd.read_csv("csv/population_final.csv")
df_internet = pd.read_csv("csv/internet_final.csv")
df_isec = pd.read_csv("csv/nb_inscrits_secondaire_final.csv")
df_isup = pd.read_csv("csv/nb_inscrits_superieur_adjusted_final.csv")
df_pour = pd.read_csv("csv/poursuite_etudes_sup_adjusted_final.csv")
df_proj = pd.read_csv("csv/projection_jeunes_diplomes_secondaire_final.csv")
```

III – Visualisation et choix pays

Deuxième notebook : P2_02_exploration

	Country Name	Country Code	Indicator Name	2016
0	United States	USA	Population, ages 15-64, total	213071223.0
1	Russian Federation	RUS	Population, ages 15-64, total	99477057.0
2	Japan	JPN	Population, ages 15-64, total	76831284.0
3	Germany	DEU	Population, ages 15-64, total	54263836.0
4	United Kingdom	GBR	Population, ages 15-64, total	42028042.0
5	France	FRA	Population, ages 15-64, total	41796373.0
6	Korea, Rep.	KOR	Population, ages 15-64, total	37364822.0

	Country Name	Indicator Name	2015
0	United States	Enrolment in tertiary education, all programmes, both sexes (number)	19531728.00
1	Russian Federation	Enrolment in tertiary education, all programmes, both sexes (number)	6592416.00
2	Japan	Enrolment in tertiary education, all programmes, both sexes (number)	3862460.00
3	Korea, Rep.	Enrolment in tertiary education, all programmes, both sexes (number)	3268099.00
4	Germany	Enrolment in tertiary education, all programmes, both sexes (number)	2977781.00
5	France	Enrolment in tertiary education, all programmes, both sexes (number)	2388880.00
6	United Kingdom	Enrolment in tertiary education, all programmes, both sexes (number)	2352932.75

III – Visualisation et choix pays

1) Matrice de corrélation

Transposition : indicateur population

Opérations réalisées pour les 6 dataframes finaux :

1) Transposition du dataframe.
2) Récupération de la Series Pandas qui nous intéresse.
3) Création d'un nouveau dataframe

Ceci est un exemple avec l'indicateur de population.

	Country Name	Country Code	Indicator Name	2016
0	United States	USA	Population, ages 15-64, total	213071223.0
1	Russian Federation	RUS	Population, ages 15-64, total	99477057.0
2	Japan	JPN	Population, ages 15-64, total	76831284.0
3	Germany	DEU	Population, ages 15-64, total	54263836.0
4	United Kingdom	GBR	Population, ages 15-64, total	42028042.0
5	France	FRA	Population, ages 15-64, total	41796373.0
6	Korea, Rep.	KOR	Population, ages 15-64, total	37364822.0

	0	1	2	3	4	5	6
Country Name	United States	Russian Federation	Japan	Germany	United Kingdom	France	Korea, Rep.
Country Code	USA	RUS	JPN	DEU	GBR	FRA	KOR
Indicator Name	Population, ages 15-64, total	Population, ages 15-64, total	Population, ages 15-64, total	Population, ages 15-64, total	Population, ages 15-64, total	Population, ages 15-64, total	Population, ages 15-64, total
2016	213071223.0	99477057.0	76831284.0	54263836.0	42028042.0	41796373.0	37364822.0

	Pays	Année	Population
0	United States	2016	213071223.0
1	Russian Federation	2016	99477057.0
2	Japan	2016	76831284.0
3	Germany	2016	54263836.0
4	United Kingdom	2016	42028042.0
5	France	2016	41796373.0
6	Korea, Rep.	2016	37364822.0

III – Visualisation et choix pays

1) Matrice de corrélation

	Pays	Population	Internet	Inscrits Enseignement Supérieur	Inscrits Enseignement Secondaire	Poursuite tertiaire	Projection : 20-39 ans tertiaire
0	United States	213071223.0	76.176737	19531728.0	11736315.0	85.795776	0.38
1	Russian Federation	99477057.0	76.409085	6592416.0	2823004.0	80.394081	0.24
2	Japan	76831284.0	92.0	3862460.0	3682920.0	63.362591	0.61
3	Germany	54263836.0	89.647101	2977781.0	2579952.25	68.265587	0.38
4	United Kingdom	42028042.0	94.775801	2352932.75	4195081.5	56.476768	0.38
5	France	41796373.0	85.6222	2388880.0	2598357.0	64.390472	0.43
6	Korea, Rep.	37364822.0	92.716545	3268099.0	1903857.0	93.179138	0.61

III – Visualisation et choix pays

1) Matrice de corrélation : interprétation

Nous observons les corrélations suivantes :

- corrélation positive de 0.98 entre le nombre d'étudiants dans le supérieur et la population : cohérent.
- corrélation positive de 0.92 entre le nombre d'étudiants dans le secondaire et la population : cohérent.
- corrélation positive de 0.94 entre le nombre d'inscrits dans le supérieur et dans le secondaire : cohérent.
- corrélation négative de -0.73 entre internet et le nombre d'inscrits dans le supérieur.
- corrélation négative de -0.77 entre internet et la population.

III – Visualisation et choix pays

1) Matrice de corrélation

Conclusion :

La matrice de corrélation n'est pas exploitable car certaines données avaient été imputées.

Comparaison de données d'années différentes : population 2016 ou 2015, internet 2016 mais les inscrits dans l'enseignement supérieur 2014 ou 2013, les projections 2020.

Ainsi, nous ne conclurons pas à partir d'une matrice de corrélation.

III – Visualisation et choix pays

2) Score d'attractivité : normalisation

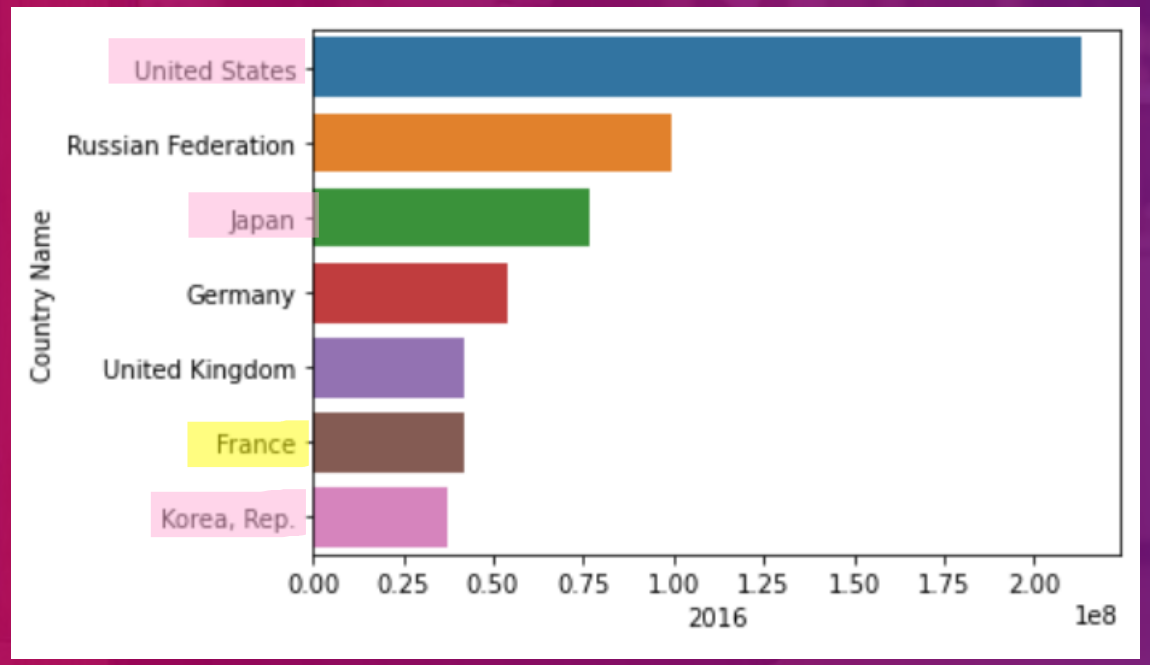
Pays	Population	Internet	Inscrits Enseignement Supérieur	Inscrits Enseignement Secondaire	Poursuite tertiaire	Projection : 20-39 ans tertiaire	Score total
United States	1.000000	0.000000	1.000000	1.000000	0.798831	0.378378	4.18
Korea, Rep.	0.000000	0.889282	0.053273	0.000000	1.000000	1.000000	2.94
Japan	0.224616	0.850756	0.087872	0.180938	0.187612	1.000000	2.53
United Kingdom	0.026540	1.000000	0.000000	0.233027	0.000000	0.378378	1.64
Germany	0.096178	0.724250	0.036373	0.068762	0.321200	0.378378	1.63
Russian Federation	0.353500	0.012492	0.246786	0.093481	0.651656	0.000000	1.36
France	0.025221	0.507846	0.002093	0.070633	0.215618	0.513514	1.33

- Nous constatons que les 6 pays retenus ont tous un meilleur score que la France que nous avions pris comme référence.
- Nous retenons les 3 meilleurs pays pour débuter le développement commercial de nos services : les Etats-Unis, la Corée du Sud et le Japon.
- Analyses univariées sur les 7 septs retenus : pour comparer pour chaque indicateur les 3 pays ayant eu le meilleur score d'attractivité.
- Puis nous verrons le potentiel d'évolution des 3 pays les mieux classés.

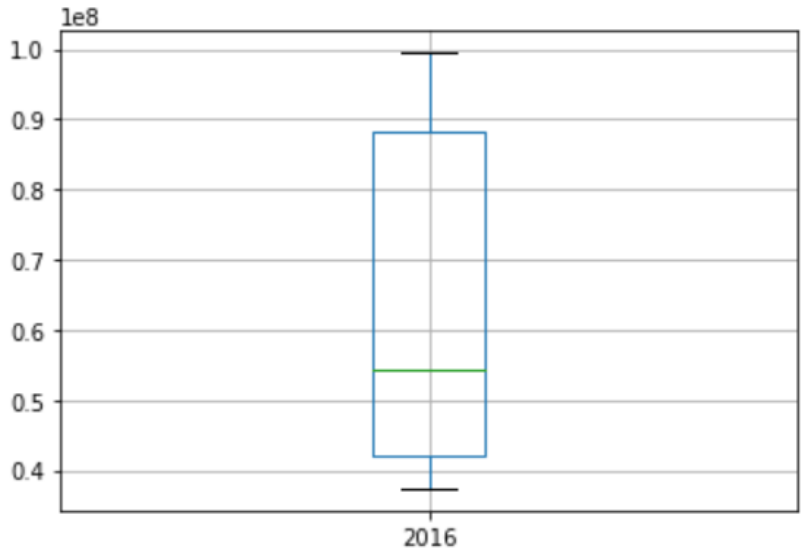
III – Visualisation et choix pays

3) Graphiques de classement

Graphique de la population des 15-64 ans par pays en 2016



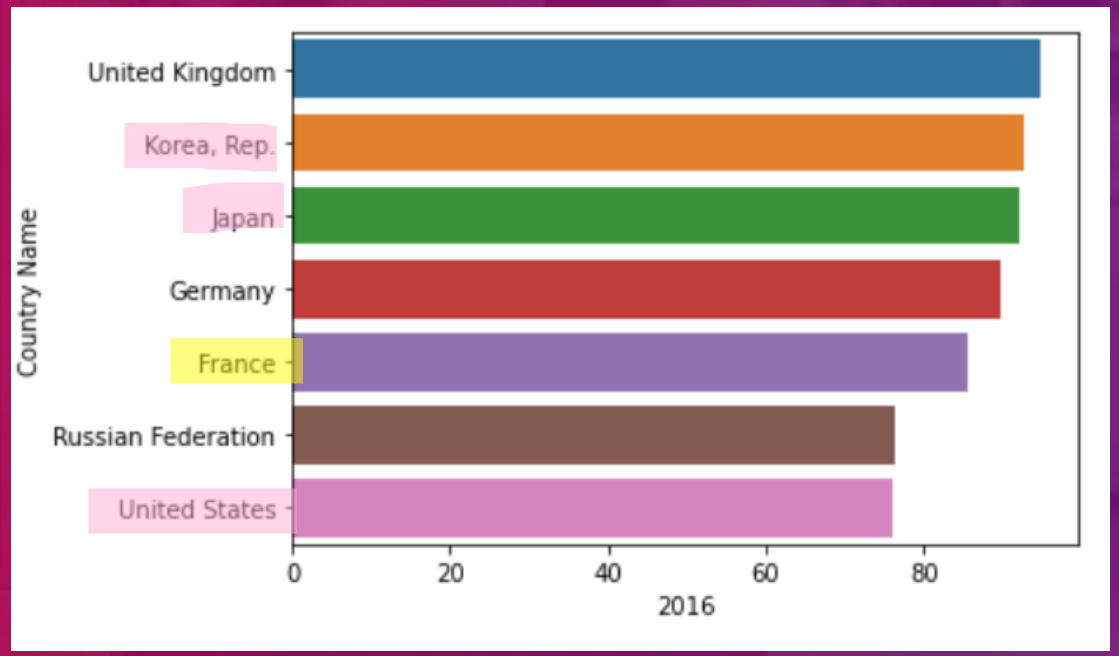

```
df_population.boxplot('2016', showfliers=False);
```

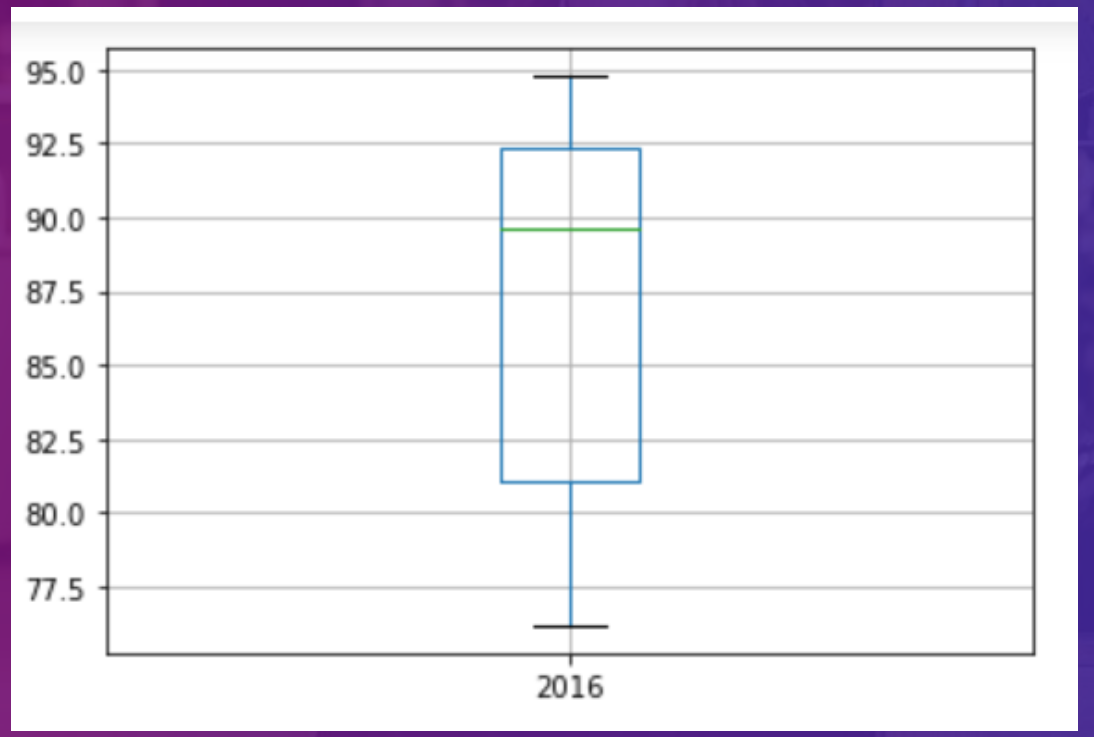


III – Visualisation et choix pays

3) Graphiques de classement

Graphique du taux d'utilisateurs d'internet (pour 100 habitants) en 2016

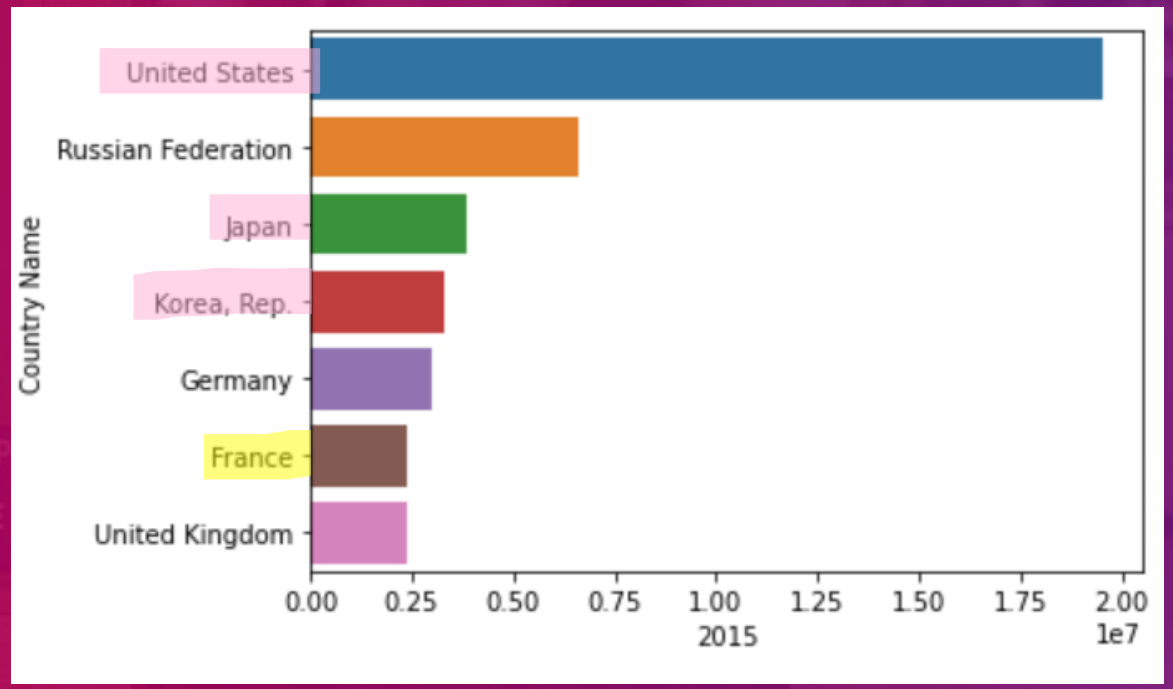




III – Visualisation et choix pays

3) Graphiques de classement

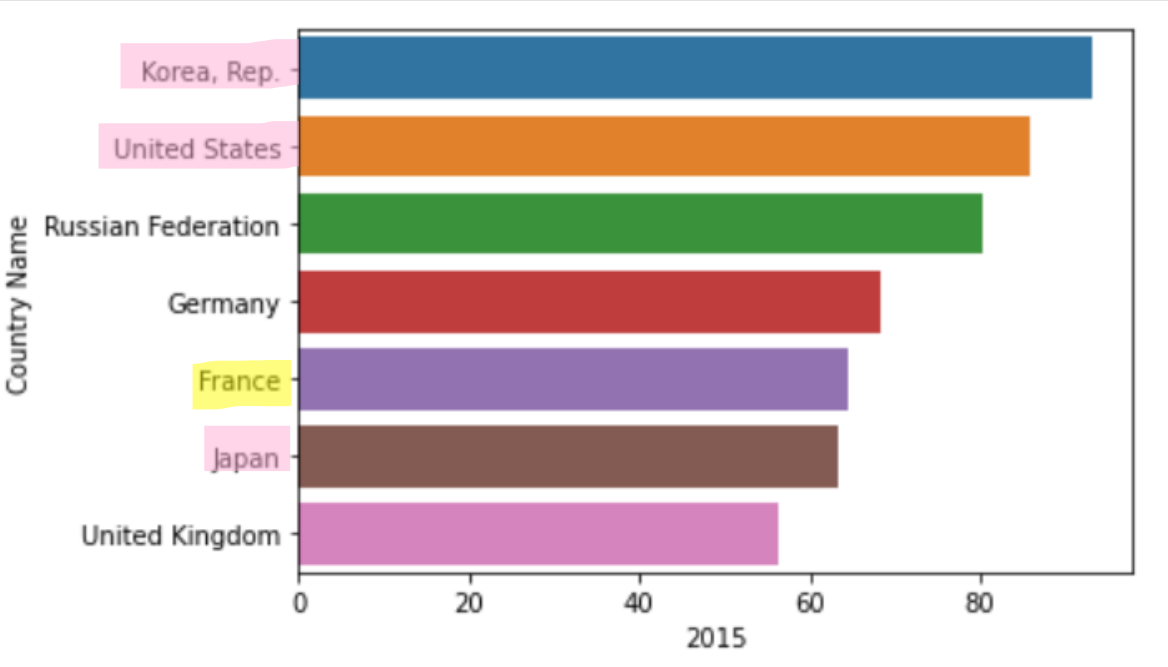
Graphique : nombre d'étudiants inscrits dans l'enseignement supérieur en 2015.



III – Visualisation et choix pays

3) Graphiques de classement

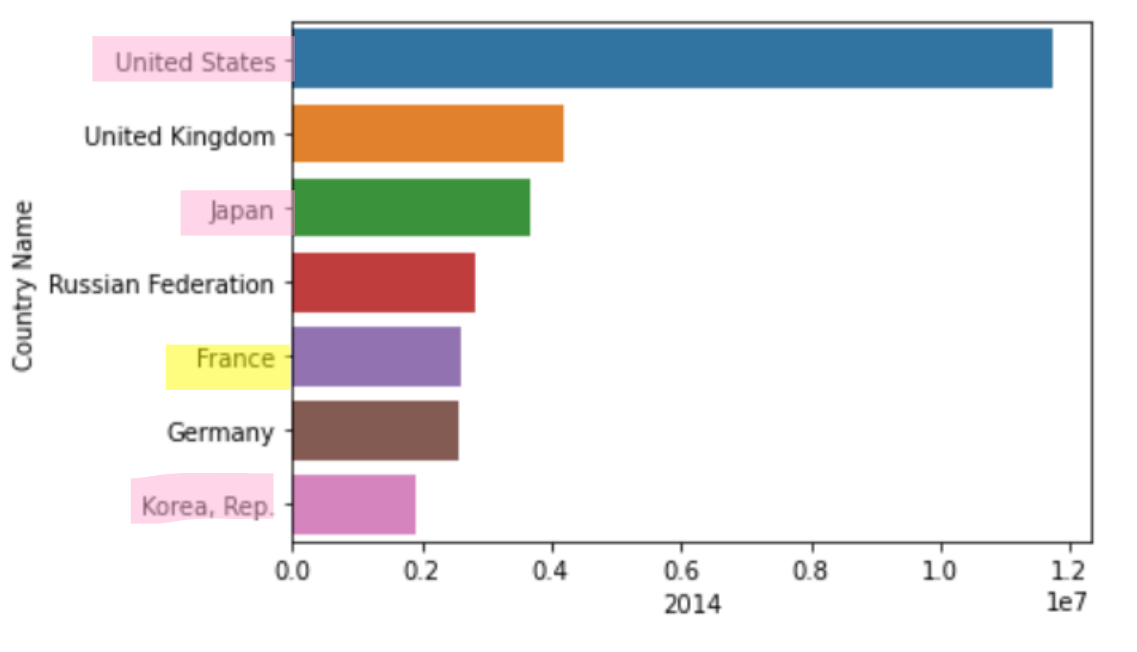
Graphique : pourcentage d'étudiants du secondaire poursuivant leurs études dans le supérieur, en 2015.



III – Visualisation et choix pays

3) Graphiques de classement

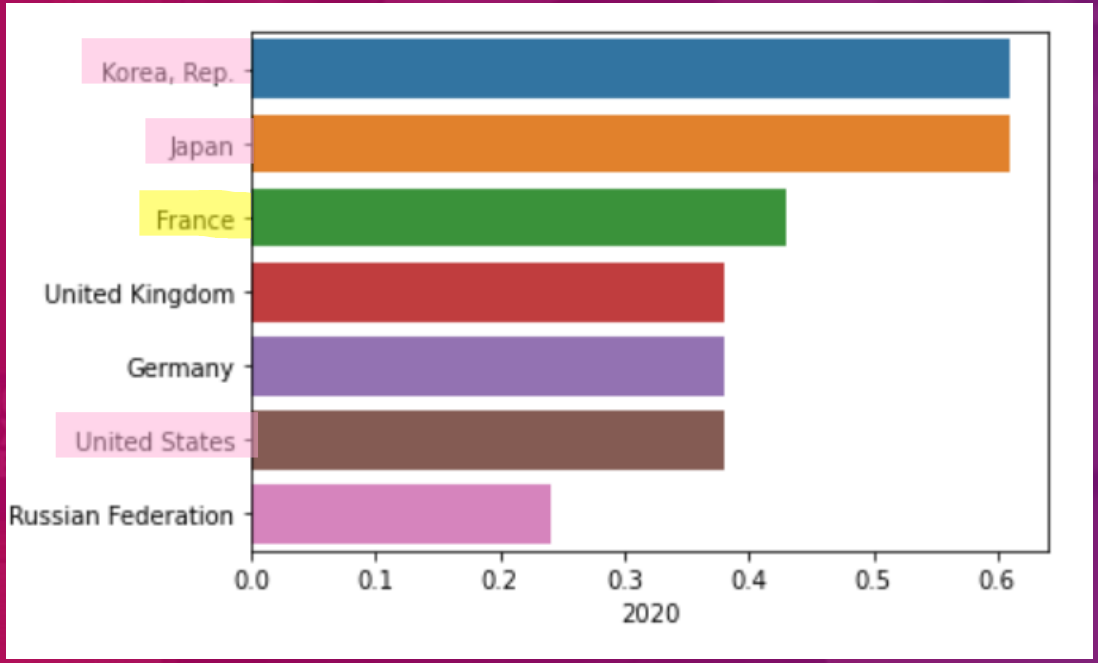
Graphique : nombre d'étudiants inscrits dans l'enseignement secondaire en 2014.

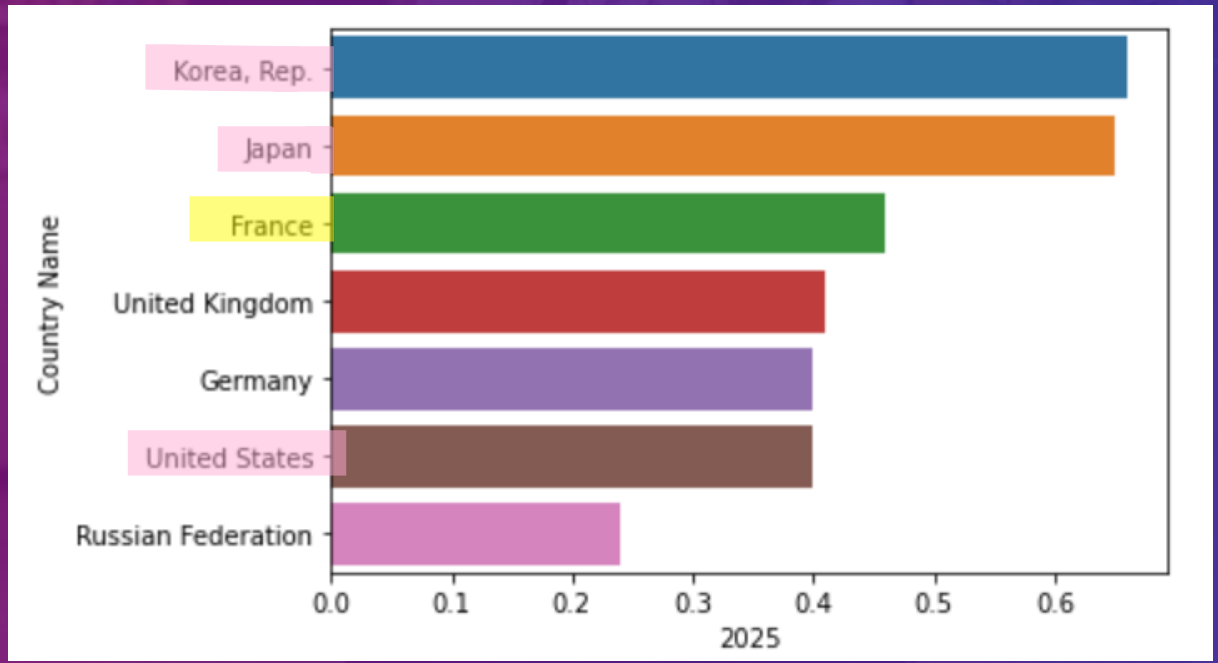


III – Visualisation et choix pays

3) Graphiques de classement

Graphiques : projection pourcentage diplômes de l'enseignement tertiaire en 2020 et 2025.



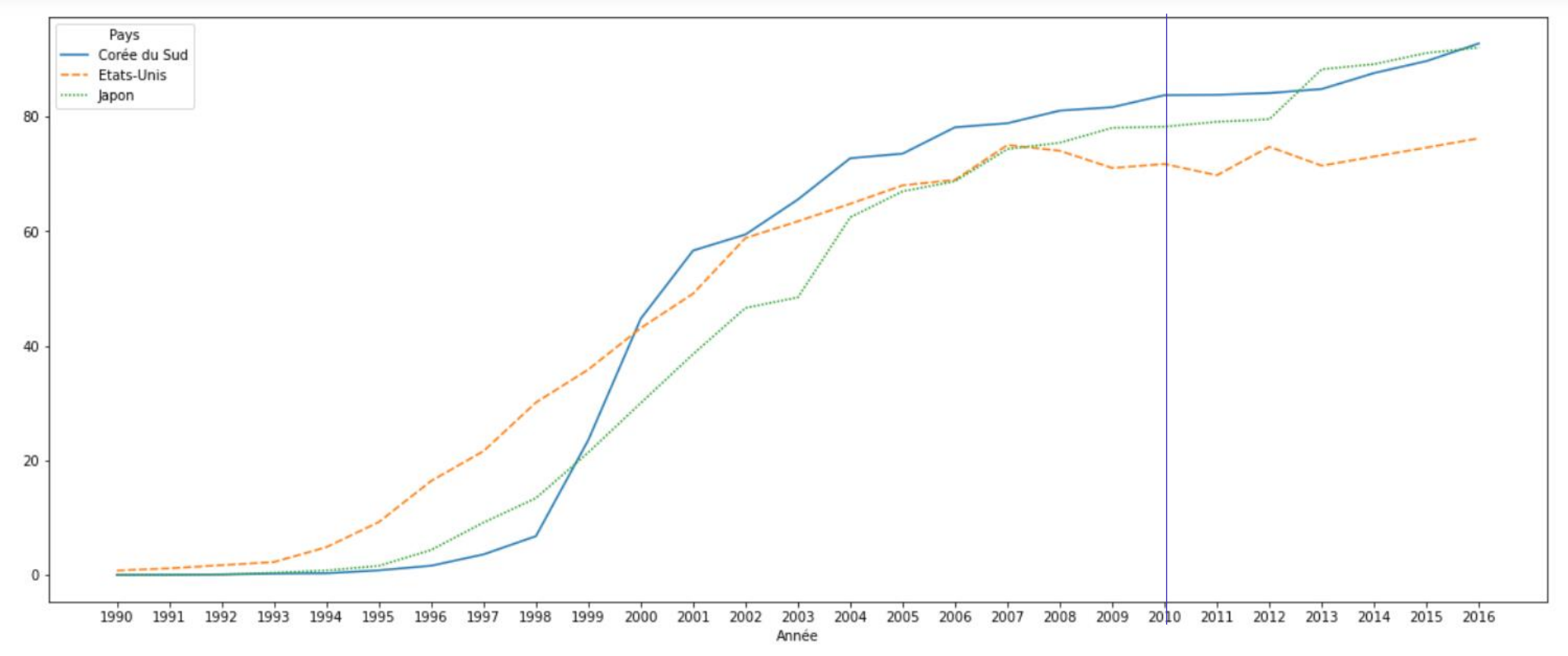


III – Visualisation et choix pays

4) Historique - internet

Table “pivot” de l’année en fonction des trois meilleurs pays pour l’utilisation d’internet (pourcentage sur 100)

Pays / Année	Corée du Sud	Etats-Unis	Japon
1990	0.023265	0.784729	0.020294
1991	0.046124	1.163194	0.040438
1992	0.098404	1.724203	0.096678
1993	0.249947	2.271673	0.401278



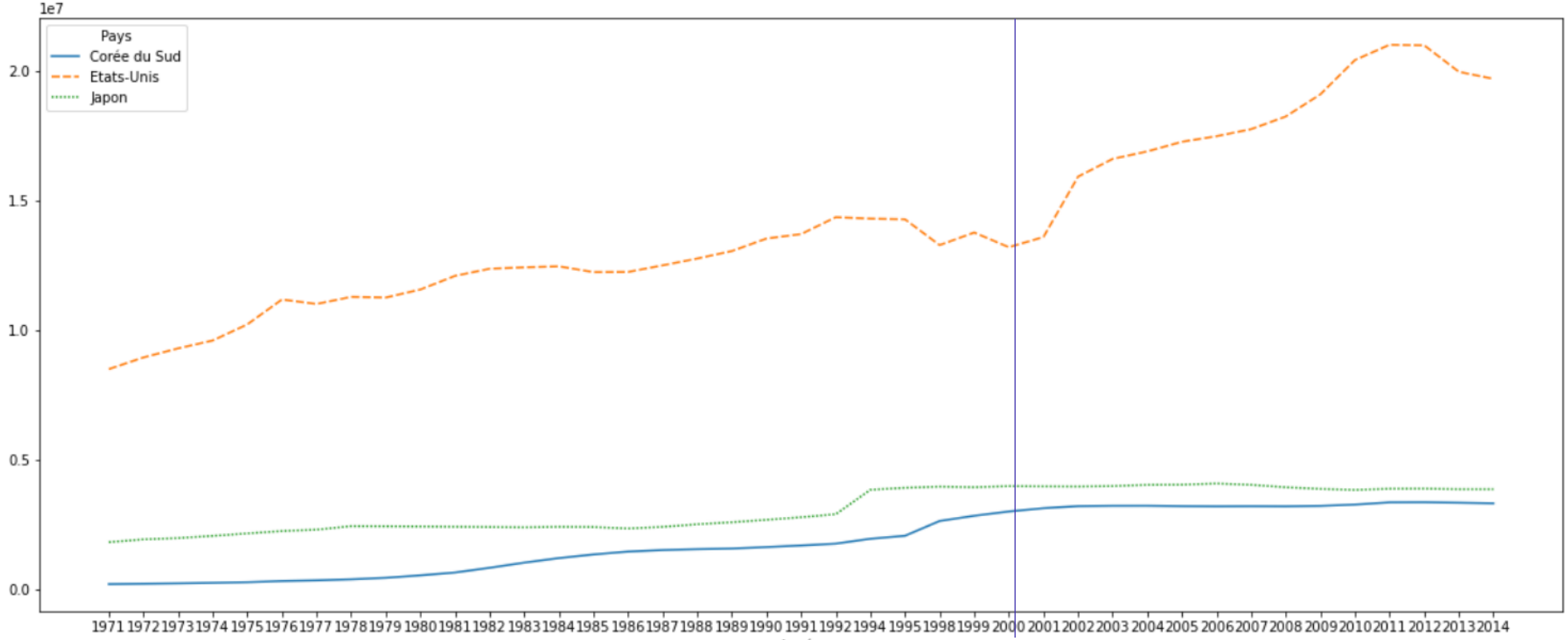
Conclusion : évolutions similaires. Japon et Corée du Sud ont un meilleur pourcentage d’utilisateurs, cependant les Etats-Unis ont une population très grande.

III – Visualisation et choix pays

4) Historique – nombre étudiants

Table "pivot" de l'année en fonction des trois meilleurs pays pour le nombre d'inscrits dans l'enseignement supérieur.

Pays Année	Corée du Sud	Etats-Unis	Japon
1971	201436.0	8498117.0	1819323.0
1972	214653.0	8948645.0	1927322.0
1973	230330.0	9297787.0	1977176.0
1974	250233.0	9602123.0	2062161.0
1975	273479.0	10223729.0	2155893.0
1976	318683.0	11184859.0	2248903.0
1977	345679.0	11012137.0	2301444.0



Conclusion :

Sur les 20 dernières années, les Etats-Unis ont un nombre beaucoup plus élevé d'étudiants que la Corée du Sud et le Japon dont les courbes sont similaires et constantes.

Conclusion

- Nous avions retenu sept pays initialement qui présentent tous d'après l'étude menée un fort potentiel pour l'implémentation et le développement des services de l'entreprise.

- L'étude à l'aide de la matrice de corrélation ne nous permet pas de conclure quand à la corrélation ou non des six indicateurs que nous avions choisis car ils sont biaisés (années différentes).

- Nous avons retenus les trois pays ayant le meilleur score : Etats-Unis, Corée du Sud et Japon.

- Sur ces trois pays, les Etats-Unis semblent être le pays présentant un potentiel de développement intéressant du fait de son nombre d'étudiants élevé.

- Le Japon et la Corée du Sud présentent une population de lycéens élevée et ont un taux de poursuite d'études supérieures élevé aussi.

Données datant au mieux de 2016 : trop anciennes pour pouvoir prédire quels sont les pays où l'entreprise devraient se développer.

- Dans l'attente de données de meilleures qualités, nous émettons l'hypothèse que l'entreprise pourrait commencer par les Etats-Unis, la Corée du Sud ou le Japon.