academy

ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

Alexandre Delaguillaumie

OBJECTIF

Réaliser une analyse exploratoire de ce jeu de données pour déterminer s'il permet d'informer ou non le projet d'expension à l'international d'academy.

PROBLÉMATIQUE

Le jeu de données permet-il d'identifier :

- Les pays avec un fort potentiel de clients
- L'évolution de ce potentiel de clients
- Les pays les plus attractifs en termes d'éducation

ROADMAP

- Présentation des données
- Exploration de données
- Nettoyage
- Sélection et construction d'indicateur
- Scoring
- Projection

I – PRÉSENTATION DES DONNÉES

I – PRÉSENTATION DES DONNÉES
DATASET 1 : DATA

Taux de remplissage
9.73%

Taille
886 930 x 70

Contenu
Pays, indicateurs, années (historiques et prospectives).

Utilité
Dataset principal pour l'exploitation des données.

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	2090	2095	2100	Unnamed: 69
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

I – PRÉSENTATION DES DONNÉES
DATASET 2 : COUNTRY

Taux de remplissage
22.23%

Taille
241 x 32

Contenu
Noms des pays, Informations génériques.

Utilité
Permet de connaître les revenus et régions associés au pays.

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	IMF data dissemination standard	Latest population census	Latest household survey	Source of most recent Income and expenditure data	Vital registration complete	agr
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD	AW	...	NaN	2010	NaN	NaN	Yes	

I – PRÉSENTATION DES DONNÉES
DATASET 3 : FOOT NOTE

Taux de remplissage
4.00%

Taille
643 638 x 4

Contenu
Informations sur l'année d'origine des données et les incertitudes.

Utilité
Si il y a un doute sur une donnée.

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN

I – PRÉSENTATION DES DONNÉES
DATASET 4 : SERIES

Taux de remplissage
5.93 %

Taille
3 665 x 21

Contenu
Toutes les informations sur les indicateurs de la banque mondiale.

Utilité
Contient des données historiques et prospectives

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

I – PRÉSENTATION DES DONNÉES
DATASET 5 : COUNTRY SERIES

Taux de remplissage
3.00%

Taille
613 x 4

Contenu
Source des données de chaque pays du dataset ‘Country’.

Utilité
Peu

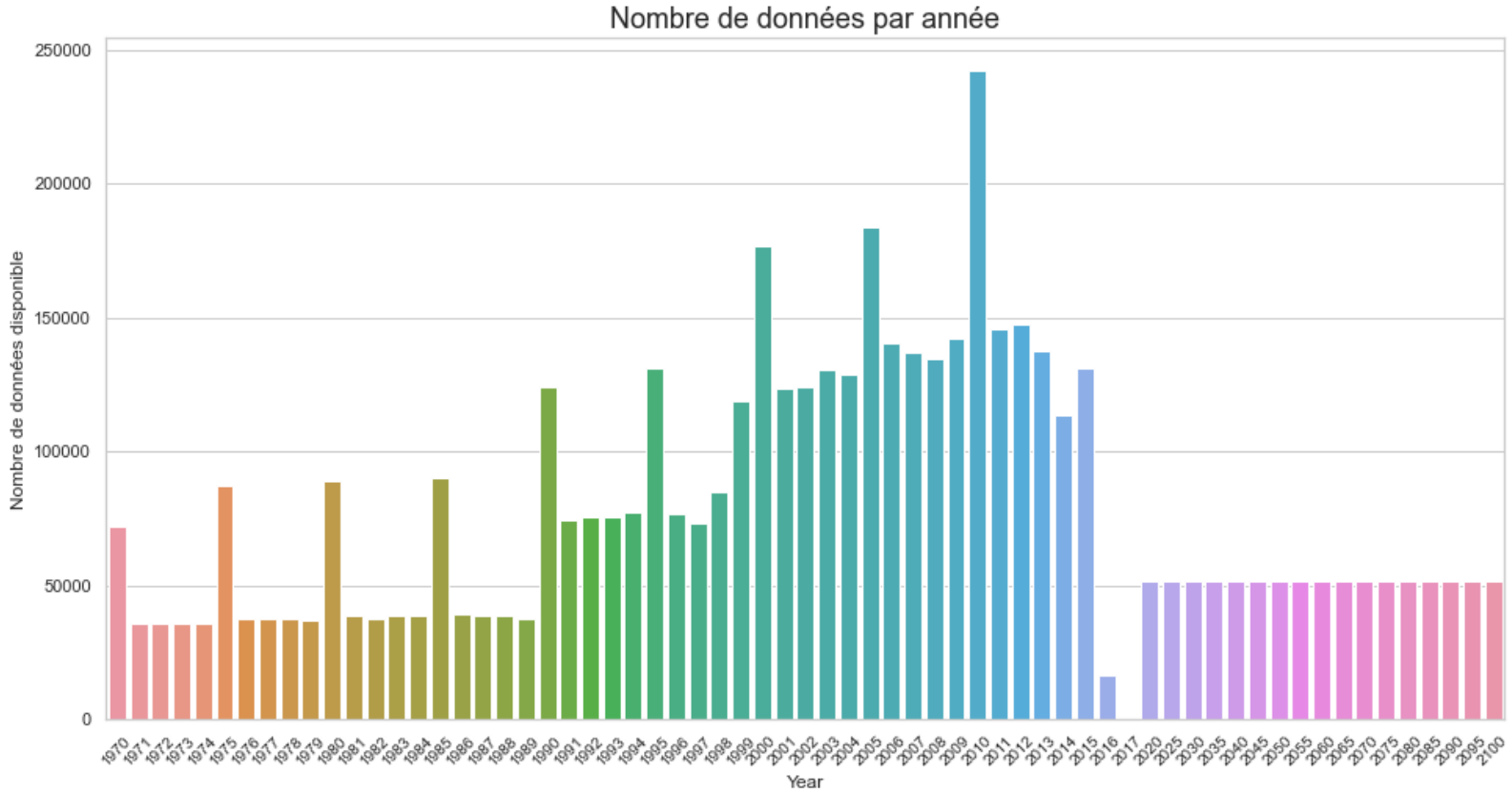
	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN

II – EXPLORATION DE DONNÉES

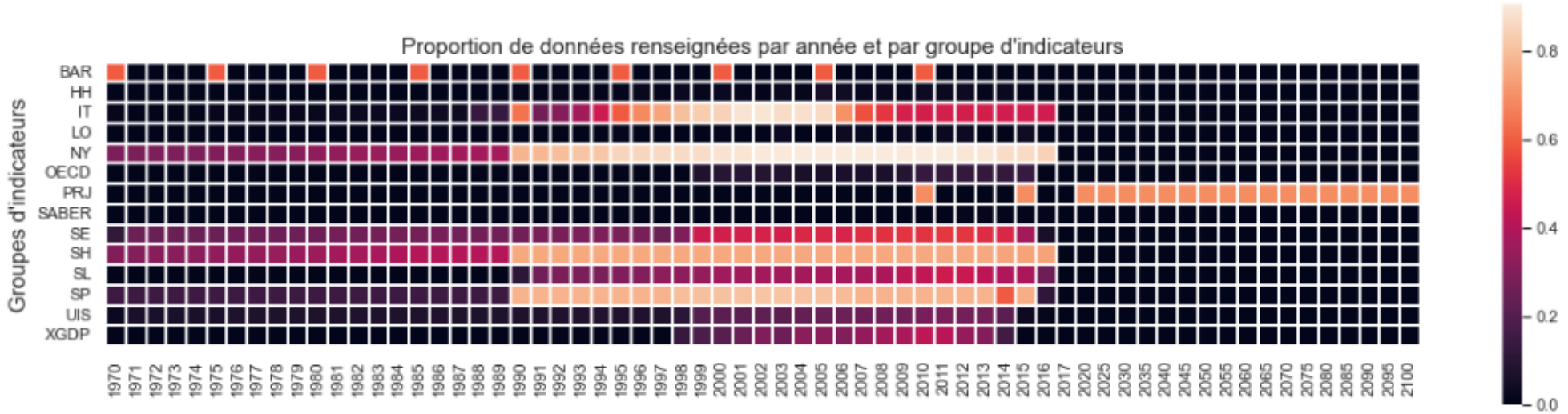
II – EXPLORATION

Indications générales
886930 x 70 colonnes
3665 indicateurs
217 pays
25 régions
65 années

Hypothèse
Décomposition des indicateurs possible.



BAR	Indicateurs Barro-Lee publiés tous les 5 ans selon 7 niveaux d'éducation.
HH	DHS (Demographic and Health Surveys) + MICS (Multiple Indicator Cluster Surveys)
IT	Infrastructure : utilisateurs internet et ordinateurs.
LO	Learning Outcomes : Evaluation du niveau des élèves (science, littérature, etc.)
NY	National Yield : balance des produits intérieurs et nationaux.
OECD	OCDE : salaires des enseignants du secteur public.
PRJ	Projections Wittgenstein (durée de scolarisation, populations, etc.)
SABER	System Approach for Better Education, aggrégation de facteurs facilitant l'accès à l'apprentissage.
SE	Social Education : comportements des différentes classes de population dans l'éducation
SL	Social Labor : mesure de la capacité à travailler des étudiants
SH	Social Health : Indicateurs de santé générale
SP	Social Population (mesure de la population selon plusieurs critères).
UIS	Unesco Institute for Statistics : ISU, données provenant de la BDD de l'UNESCO.
XGDP	Expenditure on GDP : Postes de dépenses en part de PIB (ici dépenses publiques dans l'éducation).



III – NETTOYAGE

- Exclusion des régions
- Seuil de données renseignées
- Seuil d’habitants
- Richesse minimale
- Plage d’années

III – NETTOYAGE

Exclusion des régions

Nombre de régions

25

Taille du DataFrame :

795 305

491 110

III – NETTOYAGE

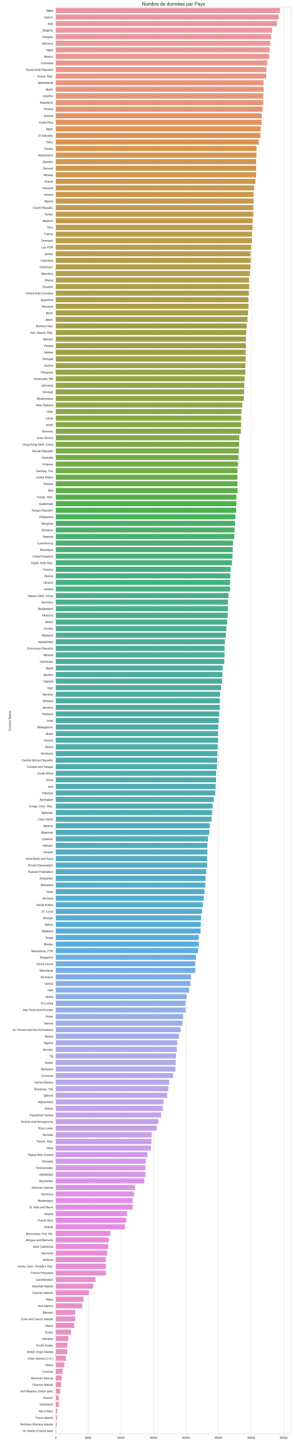
Seuil de données renseignées

Remplissage par pays
Entre 0 et 35 000

Période
De 1970 à 2100

Top / Flop
Malte, Chypre, Italie, Bulgarie, Hongrie
St-Martin, Îles Mariannes, Îles Féroé, Île de Man, Groenland

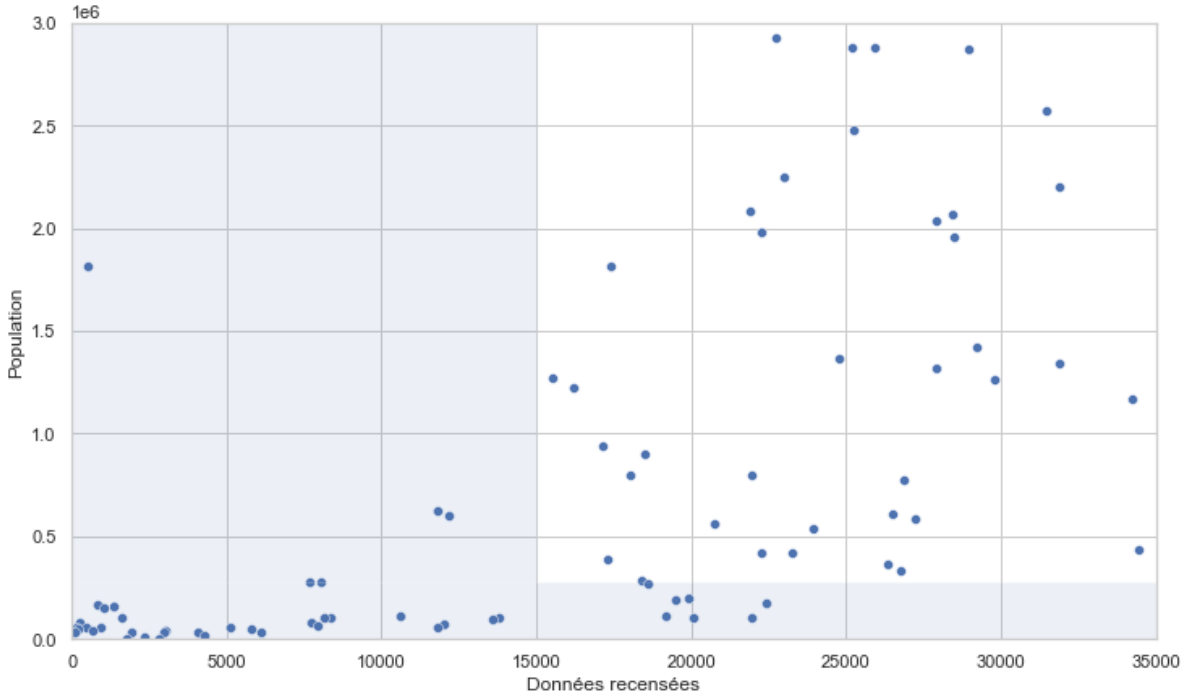
Hypothèse
Les pays recensant le moins de données sont ceux ayant le moins d'habitants et les plus pauvres.



III – NETTOYAGE
Seuil d'habitants

Comment lire le graphique
Nombre de données en fonction de la population

Filtres
280 000 habitants, 15 000 données



III – NETTOYAGE

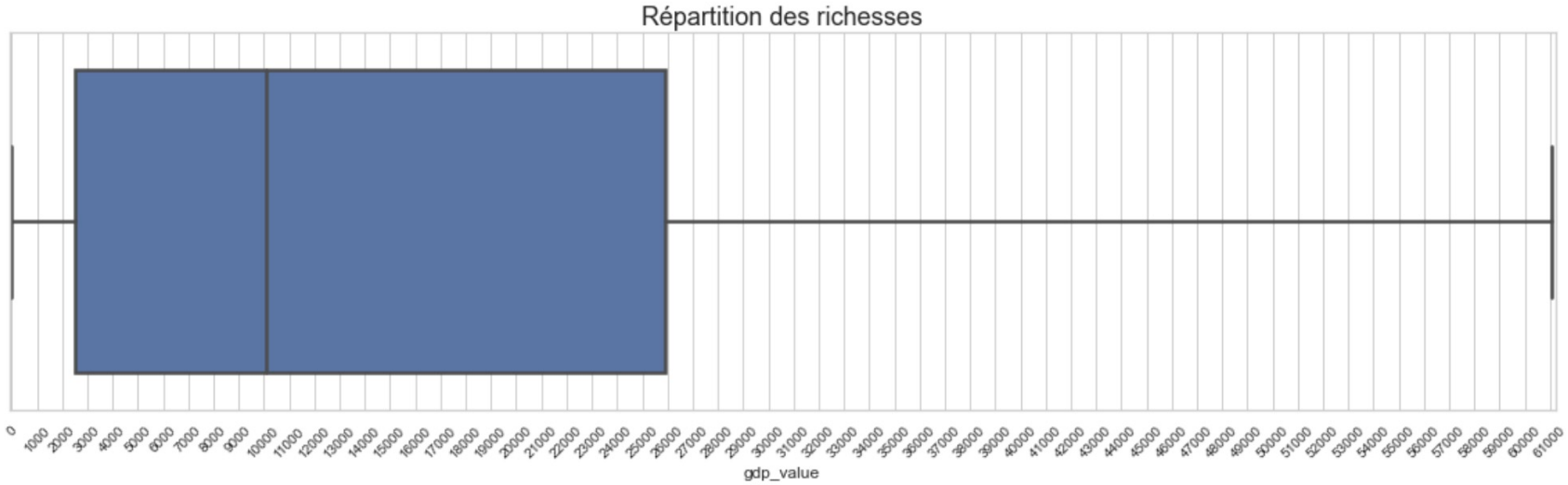
Richesse minimale

Comment lire le graphique
Répartition des richesses
par quartiles

Top / Flop
Qatar, Macao, Luxembourg, Singapore, Brunei
Centre-Afrique, Burundi, Congo, Liberia, Niger

Filtres
25 pays n'ont pas les données
de renseignées

Période
Depuis 2005

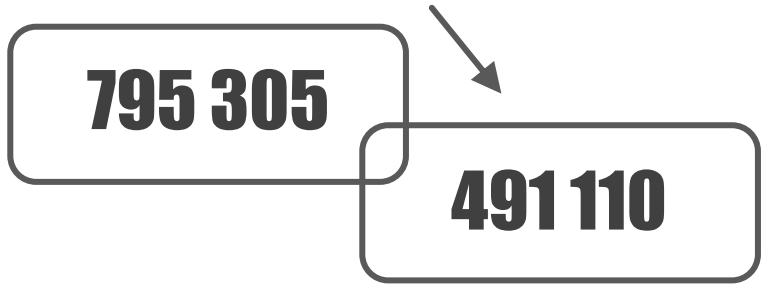


III – NETTOYAGE

Supprimer les pays de moins de :

- 15 000 données
  Seuil en dessous duquel les pays sont trop petits ou aux infrastructures insuffisantes
- 280 000 habitants
  Seuil en dessous duquel il manque beaucoup de données
- 2 482 $ de revenu par habitant
  Premier quartile

Taille du DataFrame :



IV – SELECTION

- Exclusion des groupes d’indicateurs
- Recherche sémantique
- Identifier les corrélations
- Seuil minimal de remplissage

IV – SELECTION

Exclusion des groupes d'indicateurs

BAR HH IT LO NY OECD PRJ SABER SE SL SH SP UIS XGDP

↓

BAR IT NY SE SL SP UIS XGDP

Taille du DataFrame :

491 110 → 245 220

IV – SELECTION

Recherche sémantique

Liste des mots-clés recherchés :

computer, enrolment, 8th Grade, Grade, Secondary completion, out-of-school, Population, tertiary, upper secondary, Young adults, expenditure, School census, GDP, student, Literacy, illiterate, attendance, education, technology, Numeracy, Private Sector.

Nombre d'indicateurs :

56

Taille du DataFrame :

245 220

7 236

IV – SELECTION

Étape intermédiaire

Transformation du DataFrame

Pour faire apparaître les pays groupés par région en ligne et les indicateurs en colonne

Taille du DataFrame :

7 236

134 x 56

Region	Country Name	Adult literacy rate, population 15+ years, both sexes (%)	Africa Dataset: School census return rate from secondary schools	Barro-Lee: Percentage of population age 15+ with secondary schooling. Completed Secondary	Barro-Lee: Percentage of population age 15+ with secondary schooling. Total (Incomplete and Completed Secondary)	Barro-Lee: Percentage of population age 15-19 with secondary schooling. Completed Secondary	Barro-Lee: Percentage of population age 15-19 with secondary schooling. Total (Incomplete and Completed Secondary)	Barro-Lee: Percentage of population age 15-19 with tertiary schooling. Completed Tertiary	Barro-Lee: Percentage of population age 15-19 with tertiary schooling. Total (Incomplete and Completed Tertiary)	Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary	Barro-Lee: Percentage of population age 20-24 with secondary schooling. Total (Incomplete and Completed Secondary)	...
East Asia & Pacific	Australia	NaN	NaN	38.49	60.71	32.03	79.39	0.00	0.00	76.63	96.50	...
	Brunei Darussalam	96.08556	NaN	30.69	55.28	19.14	58.02	1.28	6.62	33.41	53.18	...
	Cambodia	73.90003	NaN	4.51	17.48	1.99	25.62	0.20	1.03	8.72	25.84	...
	China	95.12448	NaN	22.85	66.47	44.43	89.63	0.65	3.38	34.49	81.07	...
	Fiji	NaN	NaN	36.33	58.87	45.34	82.26	0.43	2.20	71.26	79.26	...
...	...	...	...	...	...	...	...	...	...	...	...	...

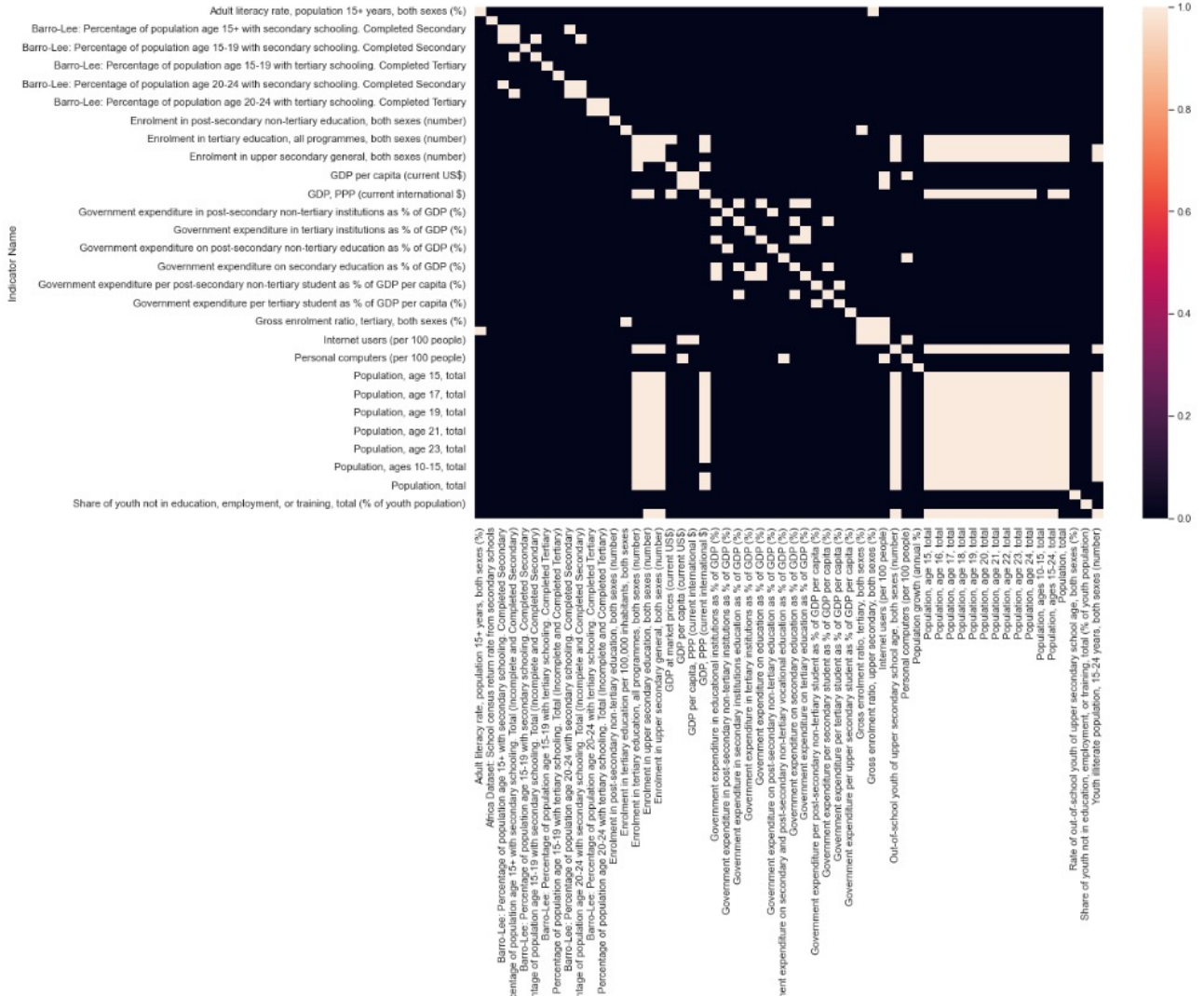
IV – SELECTION

Identifier les corrélations

Comment lire le graphique
Heatmap des indicateurs au taux de corrélation supérieur à 70%

Filtres
Sur les 56 indicateurs

Exception
Les indicateurs de population qui me seront utiles plus tard.



IV – SELECTION

Identifier les corrélations + taux de remplissage

Sélection 1 :
Manuelle de ceux qui ont les meilleurs taux de remplissage et qui répondent le mieux à la problématique

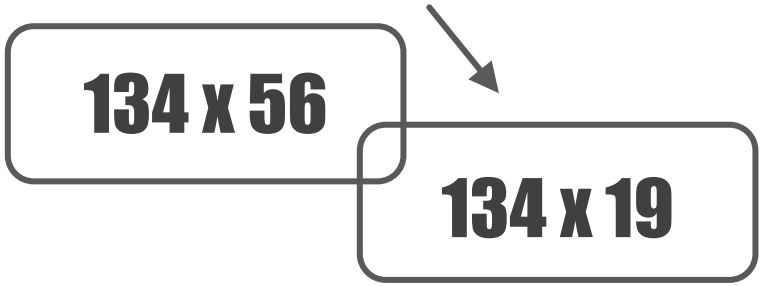
Entrée [33]:
```
correlated_indicators = [
'Government expenditure in tertiary institutions as % of GDP (%)',
'Government expenditure on tertiary education as % of GDP (%)',
'Government expenditure per tertiary student as % of GDP per capita (%)',
                        ]

for indicator in correlated_indicators:
    print(df7[indicator].isna().sum())
```

```
24
20
26
```

Sélection 2 :
Exclusion des indicateurs ayant moins de 70% de remplissage

Taille du DataFrame :



V – CONSTRUCTION

- Indicateurs de population
- Approximation du nombre d’étudiants connectés
- Scolarisation moyenne en %
- Taux de remplacement des 15-24 en %

V – CONSTRUCTION

Indicateurs de population

Pourquoi ?

Pondérer et ajuster les indicateurs existants

Lesquels ?

Population de 15 à 24 ans par année, Population, ages 10-15, total, Population, ages 15-24, total

En quoi ?

- Population ayant l'âge d'être au lycée (15-18)
  Nombre de lycéens scolarisés (taux de scolarisation appliqué aux 15-18)
- Population totale ayant l'âge d'être en études supérieures (19-24)
  Nombre d'étudiants du supérieurs scolarisés (taux de scolarisation appliqué aux 19-24)
- Notre population cible (Lycéens + Étudiants du supérieur)

V – CONSTRUCTION

Approximation du nombre d'étudiants connectés

	Étudiants Connectés en part
count	86.000000
mean	0.008066
std	0.016496
min	0.000071
25%	0.000741
50%	0.002198
75%	0.008977
max	0.100591

Pourquoi ?

Connaître le potentiel d'étudiants connectés à internet.

```
# Nombre d'étudiants connectés
df9['Utilisateur Internet en %'] = df9['Internet users (per 100 people)']
df9['Étudiants Connectés'] = df9['Utilisateur Internet en %']/100*df9['Population Étudiante']
df9['Étudiants Connectés en part'] = df9['Étudiants Connectés']/df9['Étudiants Connectés'].sum()
```

V – CONSTRUCTION

Scolarisation moyenne en %

	Scolarisation Moyenne en %
count	86.000000
mean	1.319375
std	0.384066
min	0.632068
25%	1.002516
50%	1.353141
75%	1.555168
max	2.328662

Pourquoi ?

Cumuler le nombre d'étudiants scolarisés au lycée et dans le supérieur, puis transformer ce chiffre en %.

Comment ?

Pour cela, je pondère la tranche 15-18 de 6/4 pour égaliser le poids avec la tranche 19-24, puis je fais la moyenne des 2 pourcentages.

```
# Moyenne de scolarisation des 15-24 ans
df9['Scolarisation Moyenne en %'] = (df9['Étudiants du Supérieur']/df9['19-24'])+(6/4*df9['Lycéens']/df9['15-18'])/2
```

V – CONSTRUCTION

Taux de remplacement des 15-24 en %

	Taux de remplacement des 15-24 ans
count	86.000000
mean	2.141560
std	11.642075
min	-35.838422
25%	-4.274996
50%	1.866891
75%	10.598856
max	28.440185

Pourquoi ?

Il s'agit d'avoir le pourcentage de remplacement de la population 15-24 par la population 10-15 ans. Utiliser cette méthode permet de ne pas avoir à estimer par régression linéaire l'évolution de la population 15-24 puisqu'on possède déjà les données pour le savoir.

Comment ?

Reste à ré-équilibrer les tranches d'âge puisqu'une d'elle est étalée sur 5 ans et l'autre sur 9.
On va donc appliquer 9/5 à la tranche la plus faible.

```
# Taux de remplacement des 15-24 ans
df9['Taux de remplacement des 15-24 ans'] = 100*(((9/5)*df9['Population, ages 10-15, total'])-df9['Population, ages 15-24, total'])/df9['Population, ages 15-24, total']
```

V – CONSTRUCTION

Étape intermédiaire

Renommer les 2 autres indicateurs

Pour gagner en lisibilité sur le DataFrame

	Investissement public dans l'éducation en % du PIB
count	86.000000
mean	4.913601
std	1.443693
min	1.099720
25%	3.999480
50%	4.953795
75%	5.626392
max	8.627110

	PIB par habitant en PPA
count	86.000000
mean	28390.788778
std	20523.331697
min	2985.094095
25%	11679.892763
50%	23846.021337
75%	40861.830139
max	104343.656456

```
#PIB par habitant en PPA
df9['PIB par habitant en PPA'] = df9['GDP per capita, PPP (current international $)']

# Investissement public dans l'éducation en % du PIB
df9["Investissement public dans l'éducation en % du PIB"] = df9['Government expenditure on education as % of GDP (%)'
```

V – CONSTRUCTION

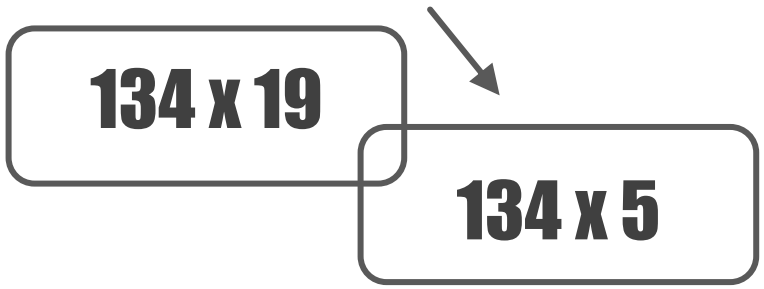
Étape intermédiaire

Transformation du DataFrame

Suppression des anciens indicateurs

Region	Country Name	Indicator Name	Étudiants Connectés en part	Scolarisation Moyenne en %	Taux de remplacement des 15-24 ans	PIB par habitant en PPA	Investissement public dans l'éducation en % du PIB
East Asia & Pacific	Australia		0.010764	2.301274	-0.937995	46789.927238	5.22534
	Brunei Darussalam		0.000103	1.003728	3.332213	77570.911947	4.42541
	Fiji		0.000107	0.732757	10.057088	9127.637483	3.88289
	Hong Kong SAR, China		0.001952	1.418779	-13.058082	58651.025580	3.26212
	Japan		0.028973	1.392846	2.194399	41476.360298	3.59184

Taille du DataFrame :



VI – FILTRAGE

Par pays

- Taux de remplissage supérieur à 80%
- Pourcentage minimal d'étudiants

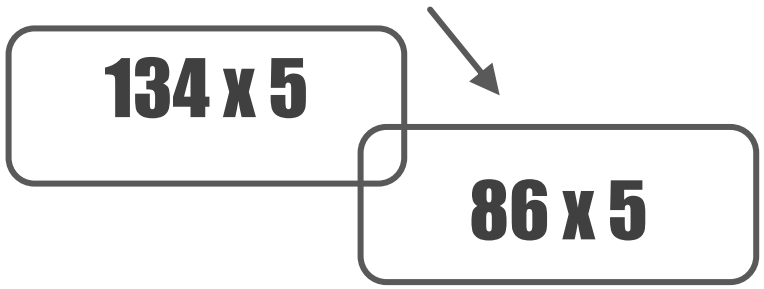
VI – FILTRAGE

Pourcentage minimal d'étudiants

Objectif :

Trouver et éliminer le quartile de pays ayant le moins d'étudiants, afin d'appliquer un filtre quantitatif de population.

Taille du DataFrame :



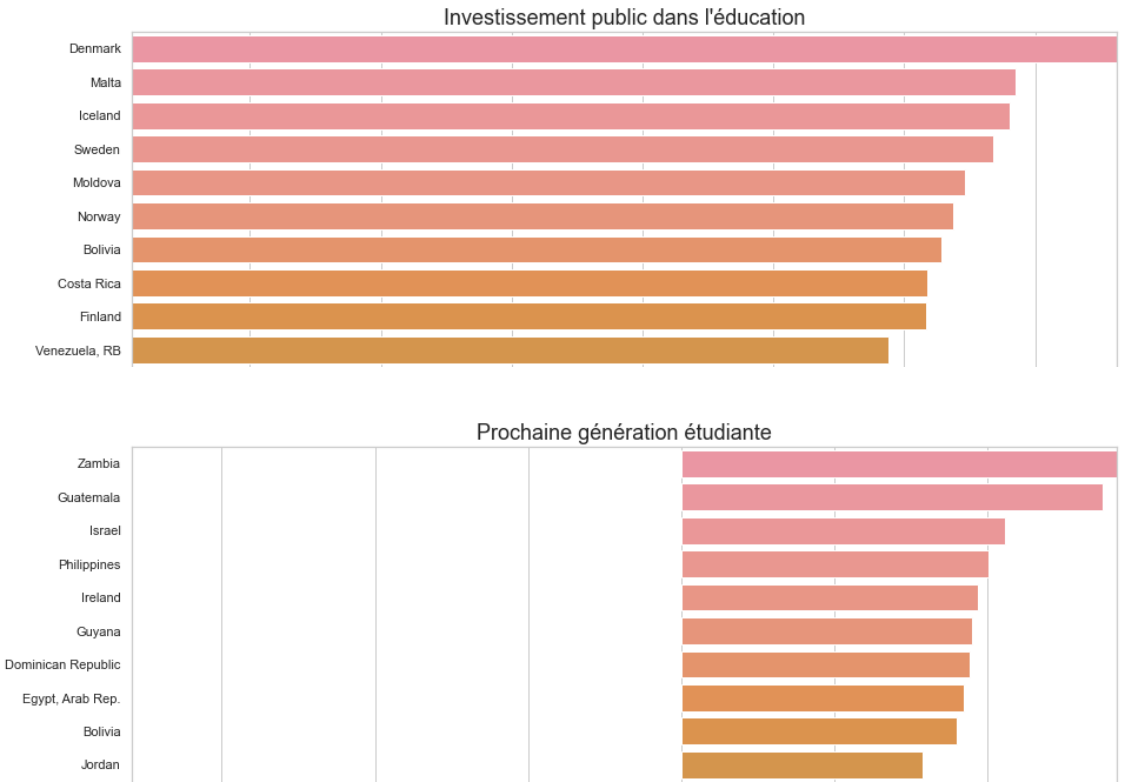

```
(df11['Population Étudiante']/df11['Population, total']*100).describe()
```

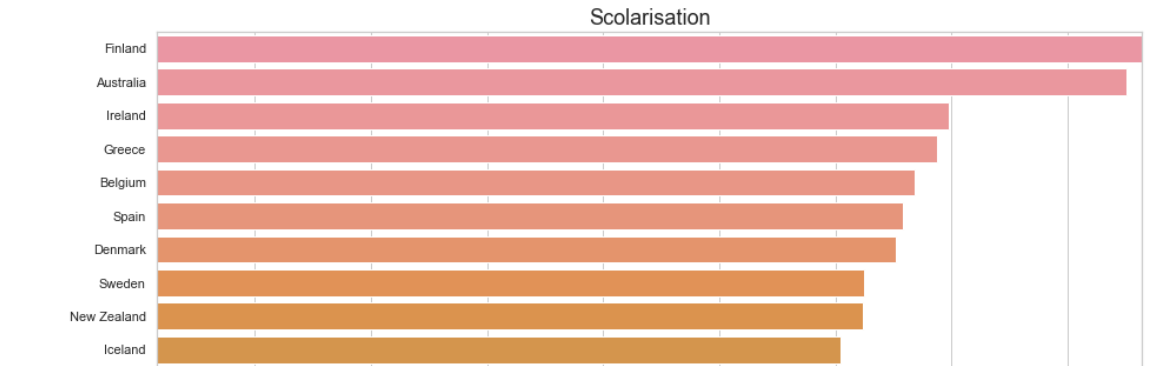
```
count    115.000000
mean       8.840238
std        3.057102
min        1.063124
25%        7.247206
50%        8.746425
75%       10.827066
max       17.397203
```

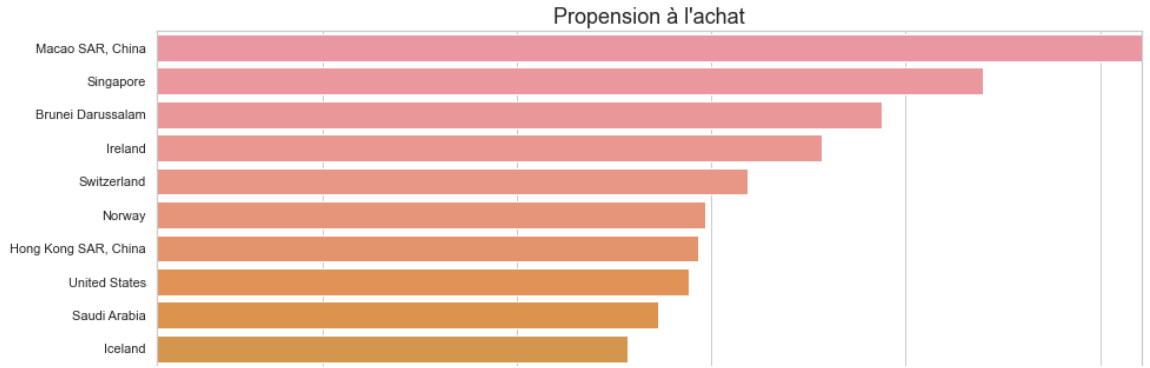
COMPARATIF DES ZONES ET PAYS

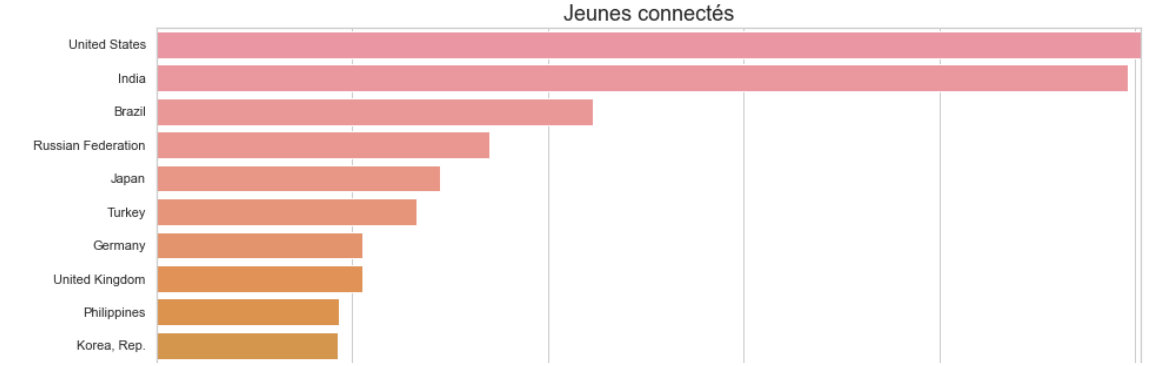
COMPARATIF

Par pays : Top 10



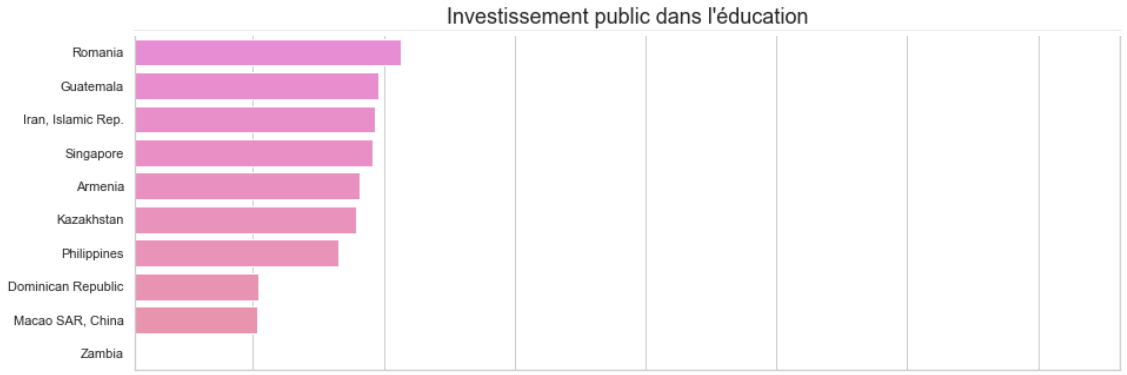


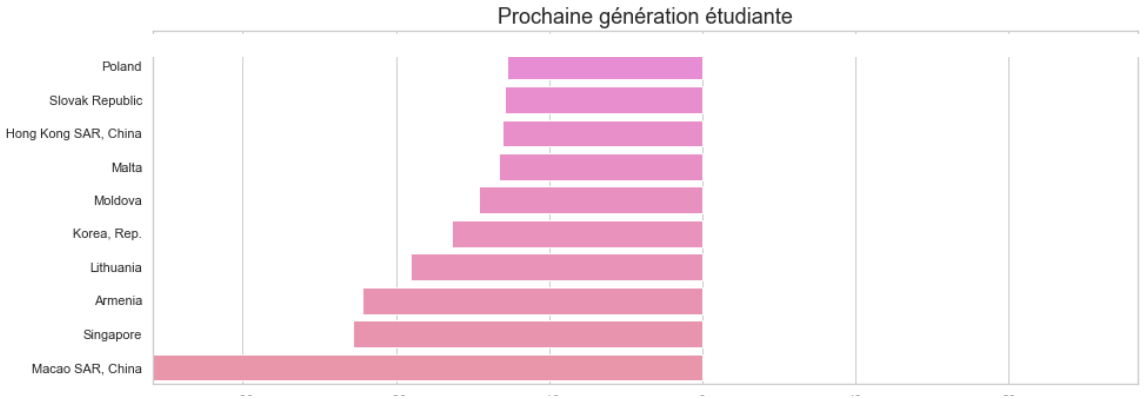


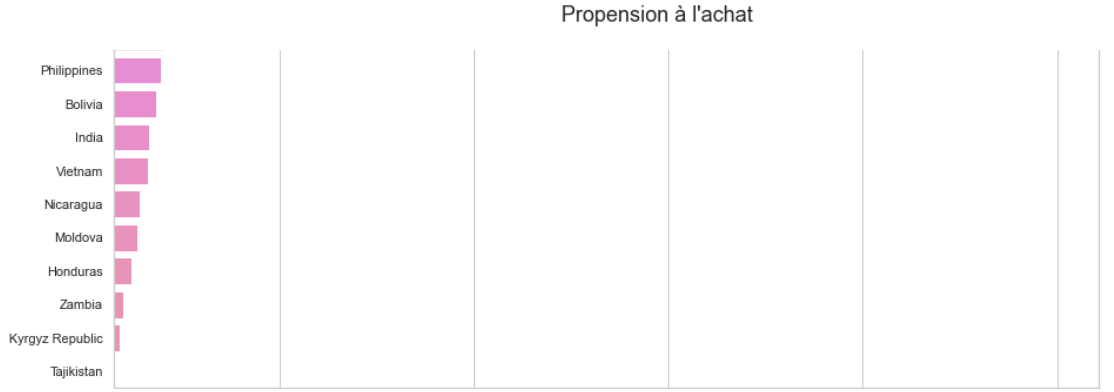


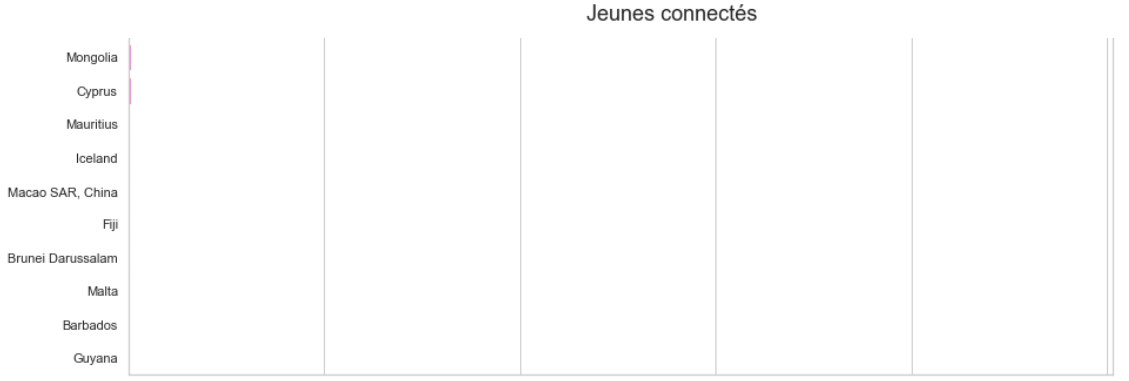
COMPARATIF

Par pays : Flop 10

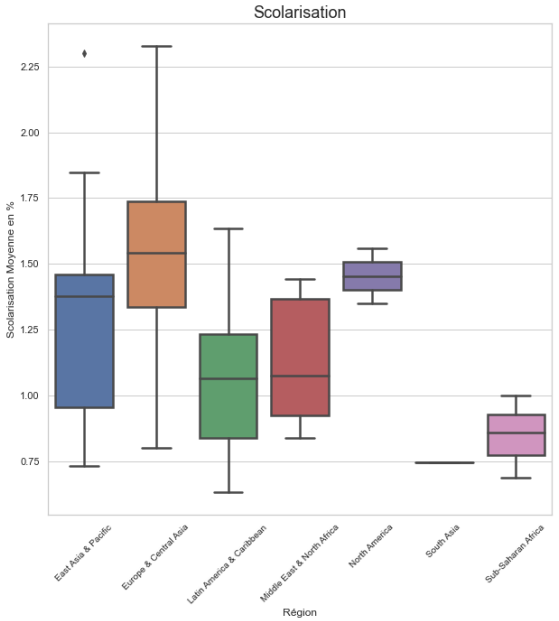


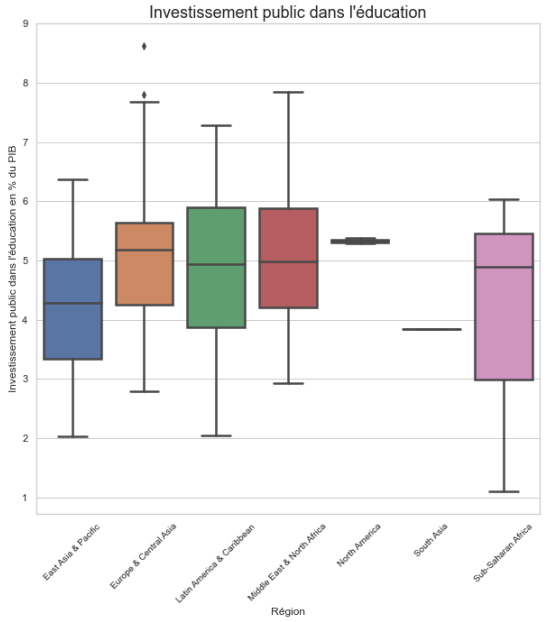


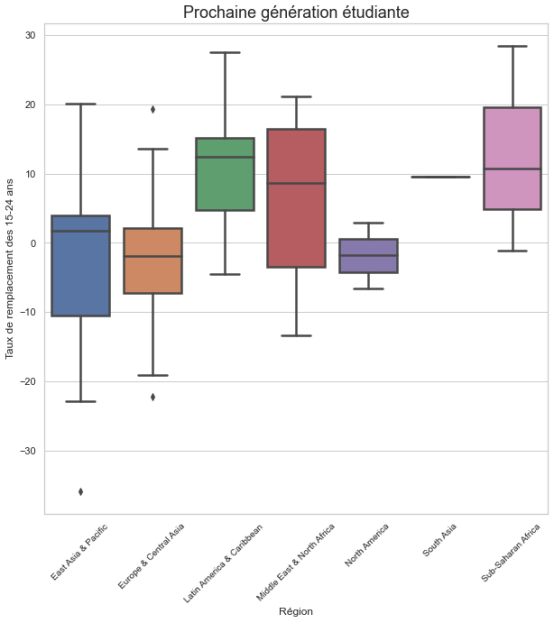


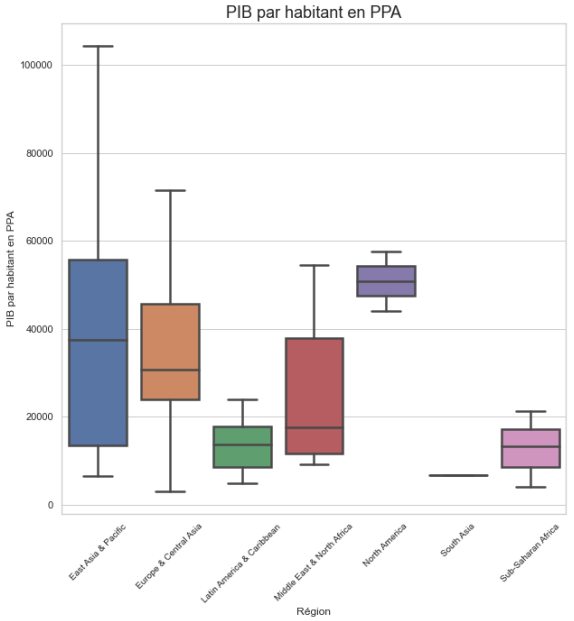


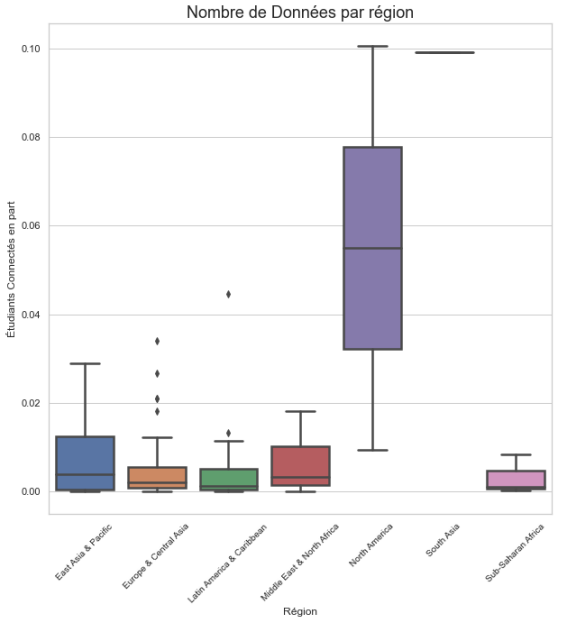
COMPARATIF
Par région











VII – SCORING

- Définition de classes
- Attribution d’un nom et de régions
- Création du score
- Merge

VII – SCORING

Définition de classes

Pourquoi ?

Le but est d'attribuer pour chaque valeur de chaque indicateur une note de 0 à 10. On découpe donc entre la valeur minimale et la valeur maximale de chaque indicateur des percentiles. Toute valeur se trouvant forcément entre 2 percentiles, une boucle conditionnelle permettra d'appliquer une note à chaque valeur.

	Étudiants Connectés en part
0.1	0.000256
0.2	0.000541
0.3	0.000988
0.4	0.001567
0.5	0.002198
0.6	0.003119
0.7	0.006606
0.8	0.010764
0.9	0.018612
1.0	0.100591

VII – SCORING

Attribution d'un nom et de régions

Pourquoi ?

afin d'attribuer plus tard le nom de l'indicateur au résultat calculé et faire un scoring par région.

	Pays	Region
0	Australia	East Asia & Pacific
1	Brunei Darussalam	East Asia & Pacific
2	Fiji	East Asia & Pacific
3	Hong Kong SAR, China	East Asia & Pacific
4	Japan	East Asia & Pacific

```
jeunes_connectes.name = 'Jeunes connectés'
pouvoir_achat.name = "Pouvoir d'achat"
scolarisation.name = 'Scolarisation'
next_gen.name = 'Renouvellement étudiant'
public_investment.name = 'Public investment'
```

VII – SCORING

Création du score

```
def ScoreMaker(indicateur,score):
    indicateur.index = indicateur['Distribution par pays']
    score_final = {}
    for pays in indicateur.index:
        if indicateur.loc[pays][2] > score.iloc[8][0]:
            score_final[pays] = 9
        elif indicateur.loc[pays][2] > score.iloc[7][0]:
            score_final[pays] = 9
        elif indicateur.loc[pays][2] > score.iloc[6][0]:
            score_final[pays] = 8
        elif indicateur.loc[pays][2] > score.iloc[5][0]:
            score_final[pays] = 7
        elif indicateur.loc[pays][2] > score.iloc[4][0]:
            score_final[pays] = 6
        elif indicateur.loc[pays][2] > score.iloc[3][0]:
            score_final[pays] = 5
        elif indicateur.loc[pays][2] > score.iloc[2][0]:
            score_final[pays] = 4
        elif indicateur.loc[pays][2] > score.iloc[1][0]:
            score_final[pays] = 3
        elif indicateur.loc[pays][2] > score.iloc[0][0]:
            score_final[pays] = 2
        else: score_final[pays] = 1


    df_score = pd.DataFrame(list(score_final.items()), columns=['Pays','Score'])
    df_score['Indicateur'] = indicateur.name
    df_fin = df_pays_region.merge(df_score, on='Pays')
    return df_fin
```

VII – SCORING

Merge

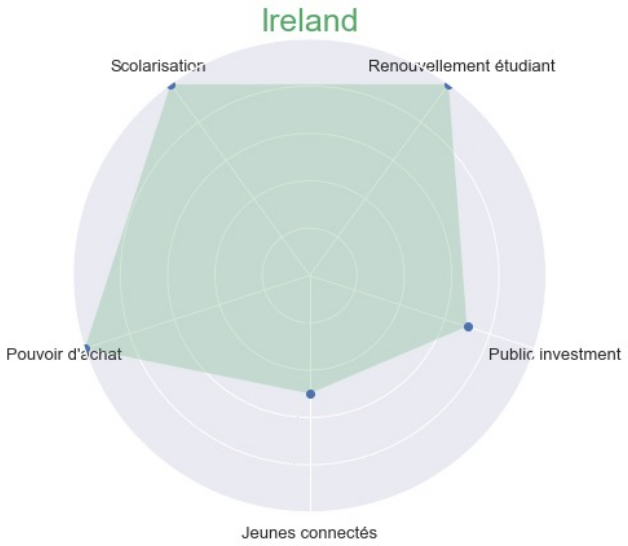
Forme du tableau
Long-form pour faciliter l'exploitation de données

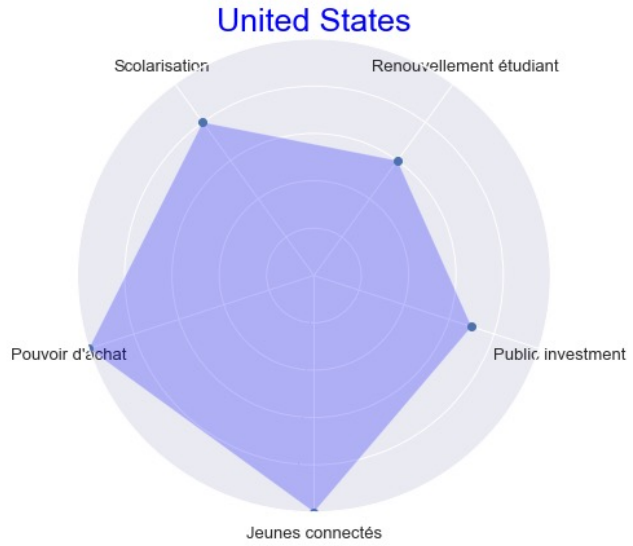
Opérations
Il ne reste plus qu'a regrouper et classer par ordre décroissant

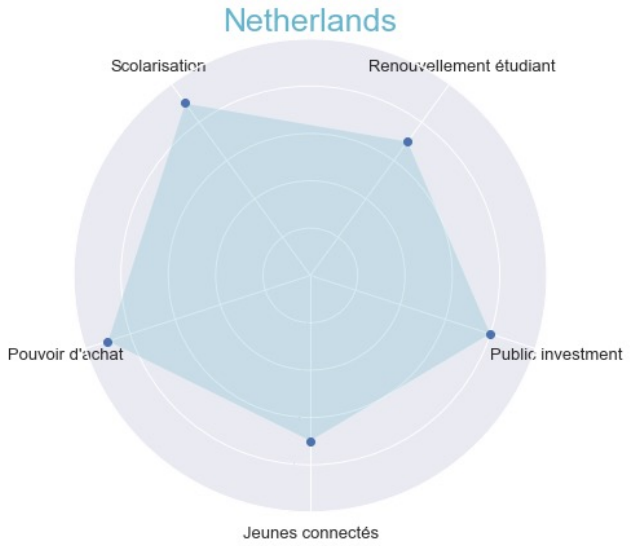
	Pays	Region	Score	Indicateur
0	Australia	East Asia & Pacific	8	Jeunes connectés
1	Brunei Darussalam	East Asia & Pacific	1	Jeunes connectés
2	Fiji	East Asia & Pacific	1	Jeunes connectés
3	Hong Kong SAR, China	East Asia & Pacific	5	Jeunes connectés
4	Japan	East Asia & Pacific	9	Jeunes connectés
...	...	...	...	...
420	United States	North America	7	Public investment
421	India	South Asia	2	Public investment
422	Mauritius	Sub-Saharan Africa	5	Public investment
423	South Africa	Sub-Saharan Africa	9	Public investment
424	Zambia	Sub-Saharan Africa	1	Public investment

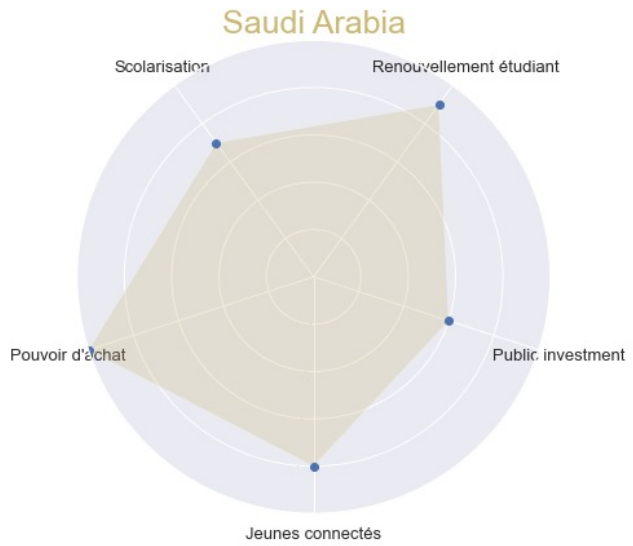
TOP PAYS ET RÉGIONS

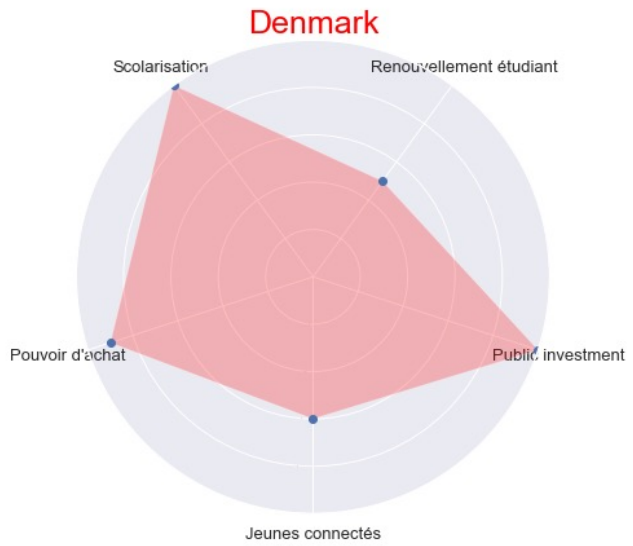
TOP 5 PAYS


Ireland
Scolarisation
Renouvellement étudiant
Pouvoir d'achat
Public investment
Jeunes connectés


United States
Scolarisation
Renouvellement étudiant
Pouvoir d'achat
Public investment
Jeunes connectés

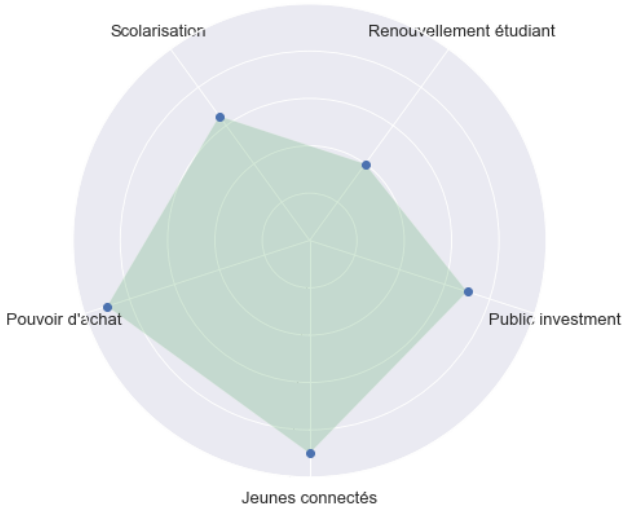

Netherlands
Scolarisation
Renouvellement étudiant
Pouvoir d'achat
Public investment
Jeunes connectés


Saudi Arabia
Scolarisation
Renouvellement étudiant
Pouvoir d'achat
Public investment
Jeunes connectés

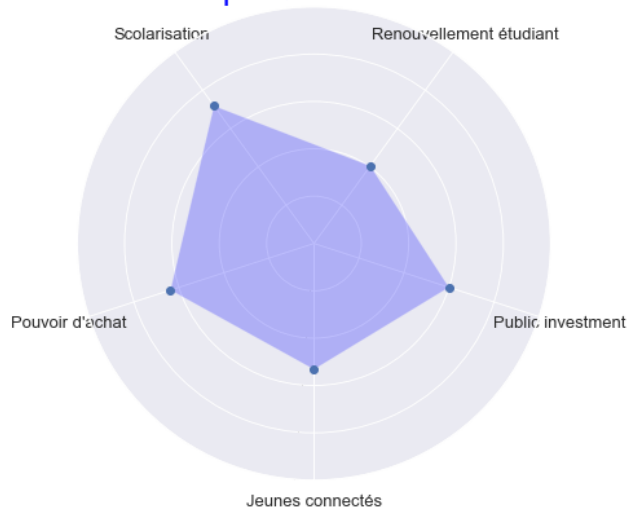

Denmark
Scolarisation
Renouvellement étudiant
Pouvoir d'achat
Public investment
Jeunes connectés

TOP 5 REGION

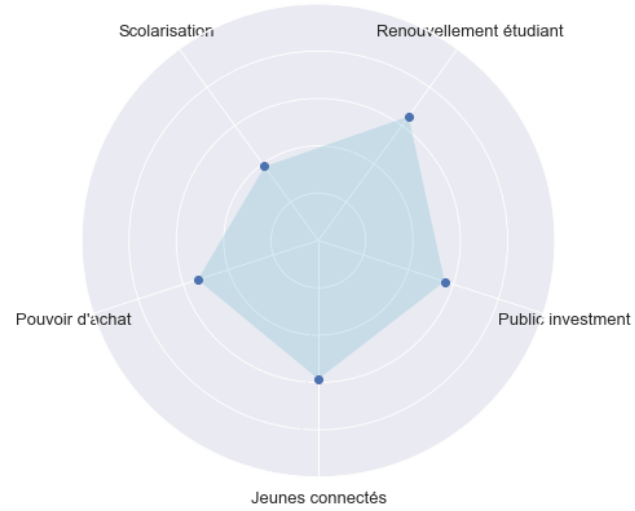
North America



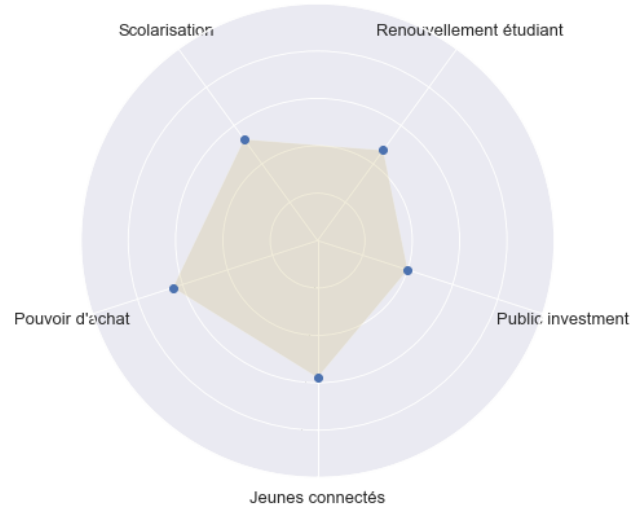
Europe & Central Asia



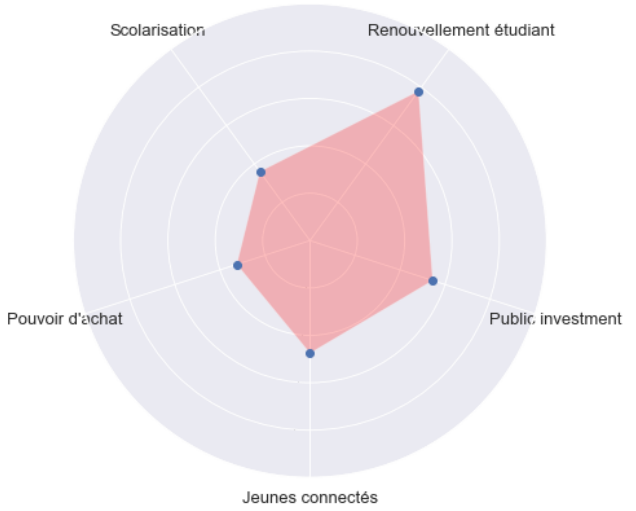
Middle East & North Africa



East Asia & Pacific



Latin America & Caribbean



VIII – PROJECTION

- Recherche sémantique
- Création d’un DataFrame dédié
- Construction d’un indicateur de projection

IV – SELECTION
Recherche sémantique

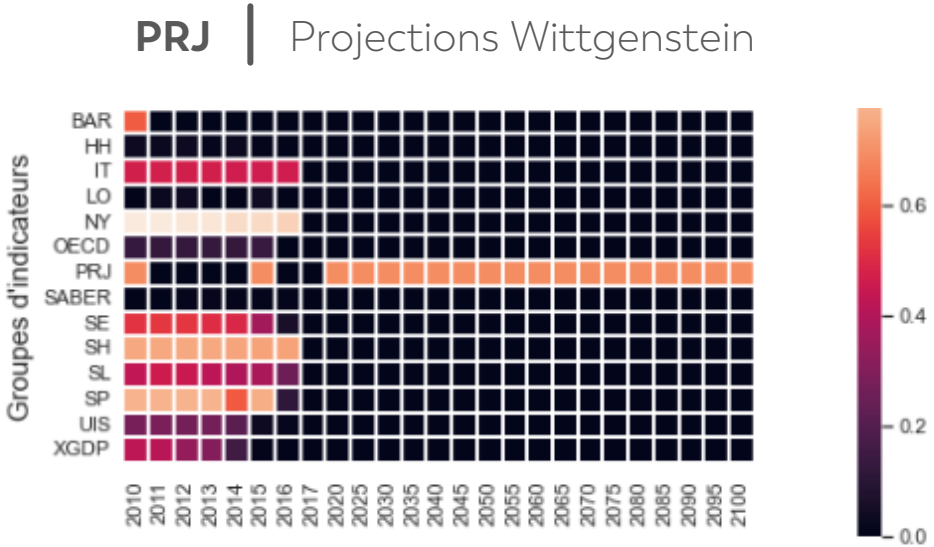
Liste des mots-clés recherchés :

Sélectionner les indicateurs de projections Wittgenstein en repartant du dernier DataFrame en date qui n'a pas fait de tri d'indicateurs.

```
indicators_in_subcategory =  projections['Indicator Name'].unique()
preselec = []

def SearchKeyword(s):
    for indicator in indicators_in_subcategory:
        if s in indicator:
            preselec.append(indicator)

SearchKeyword('Mean years of schooling.')
print(preselec)
```



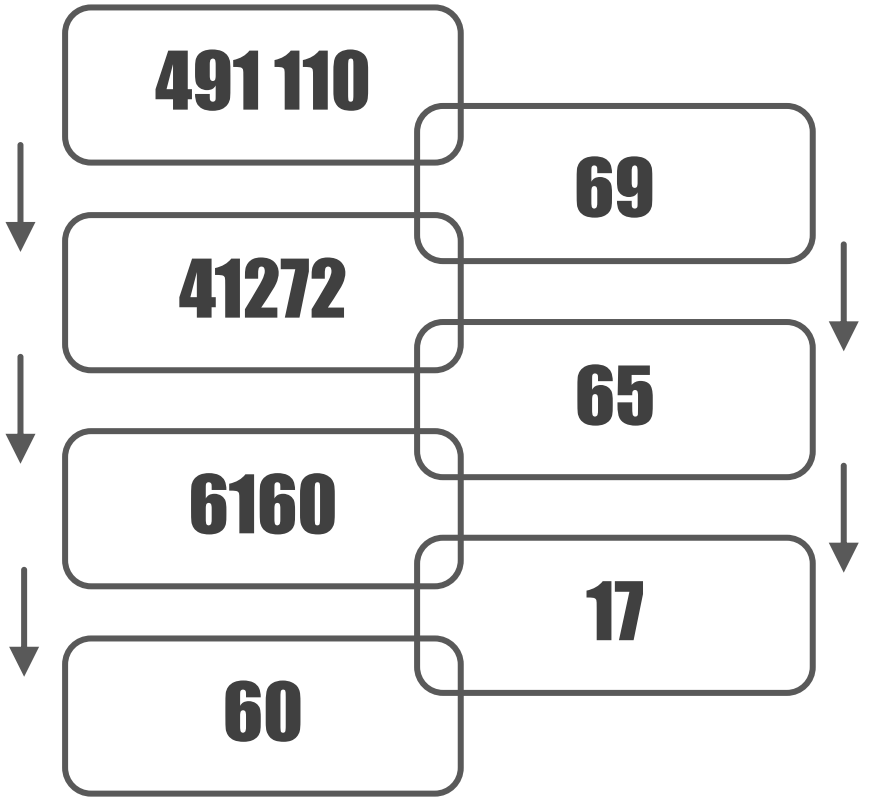
IV – SELECTION

Création d’un DataFrame dédié

Données à récupérer :

- Une plage d'année valide
- Pré-séléction d’indicateurs de projection
- Les pays du top 20

Taille du DataFrame :



IV – SELECTION

Construction d'un indicateur unique de projection

Pourquoi ?

Représenter l'évolution éducationnelle potentielle de 20 pays.

Par la fusion de 3 indicateurs sélectionnés:

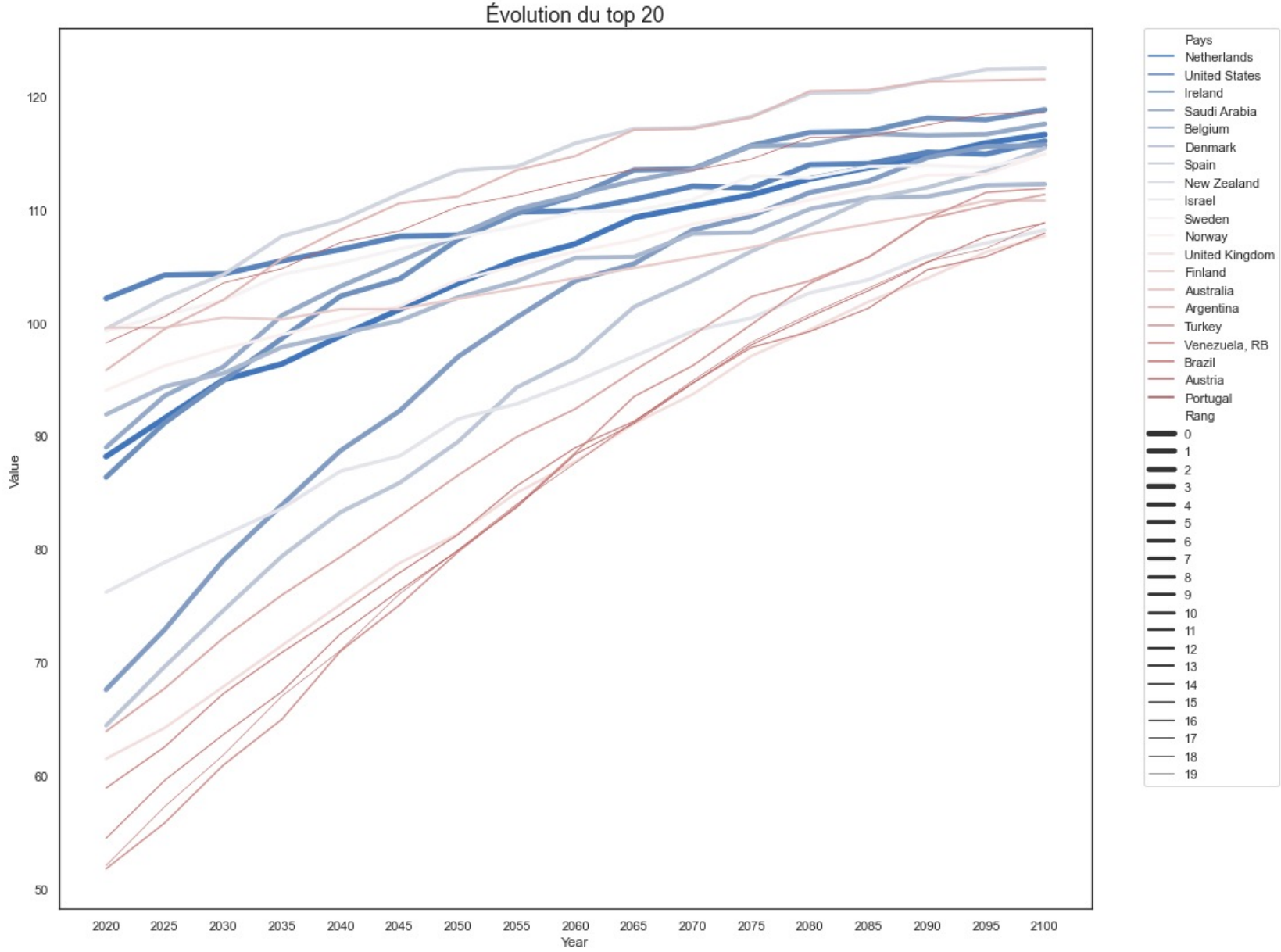
- Le nombre d'années d'études moyen
- Le pourcentage de la population ayant terminé ses études au lycée
- Le pourcentage de la population ayant terminé ses études en études supérieures

	Rang	Pays	Year	Value
0	0	Netherlands	2020	88.177476
1	0	Netherlands	2025	91.561924
2	0	Netherlands	2030	94.946371
3	0	Netherlands	2035	96.358208
4	0	Netherlands	2040	98.835408
...	...	...	...	...
335	19	Portugal	2080	100.744081
336	19	Portugal	2085	103.063165
337	19	Portugal	2090	105.382250
338	19	Portugal	2095	106.541792
339	19	Portugal	2100	108.860876

Calcul : (Va-Vd)/Vd*100

ÉVOLUTION DU TOP 20

ÉVOLUTION DU TOP 20



CONCLUSION

+

- Diversité des indicateurs
- Quantité de pays, données remplies pour les pays intéressants
- Données sourcées, détaillées, datées, non-dupliquées

-

- Manque certains indicateurs essentiels
- N’est pas actualisé
- Beaucoup d’indicateurs inutilisables

MERCI