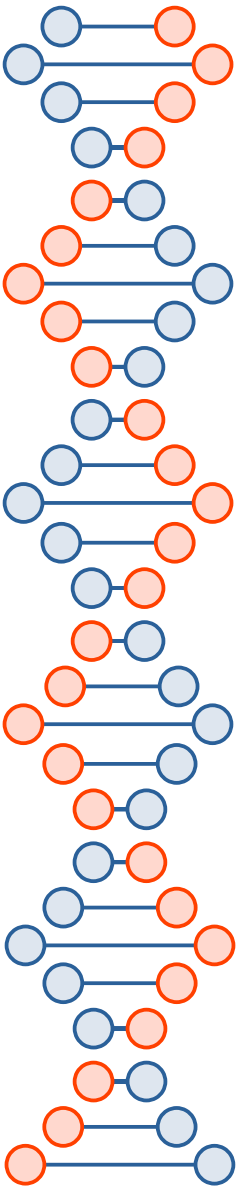


Analyse de données de systèmes éducatifs

Projet 2
Sofia Velasco



Qualité du jeu de données



Caractéristiques générales du jeu de données

- **886 930** lignes et **70** Colonnes.
- **Aucune** ligne n'est dupliquée.

Colonnes	Type de données
Country Name	object
Country Code	object
Indicator Name	object
Indicator Code	object
Années (64 colonnes suivantes)	float64

[illegible]

Étude des colonnes 'Country Name' et 'Country Code'

242 lignes telles que :

- 25 lignes 'générales' --

- 217 pays -----

	Country Name	Country Code
0	Arab World	ARB
3665	East Asia & Pacific	EAS
7330	East Asia & Pacific (excluding high income)	EAP
10995	Euro area	EMU
14660	Europe & Central Asia	ECS
18325	Europe & Central Asia (excluding high income)	ECA
87960	World	WLD
...
95290	Albania	ALB
98955	Algeria	DZA
102620	American Samoa	ASM
106285	Andorra	AND
109950	Angola	AGO
113615	Antigua and Barbuda	ATG
117280	Argentina	ARG
120945	Armenia	ARM



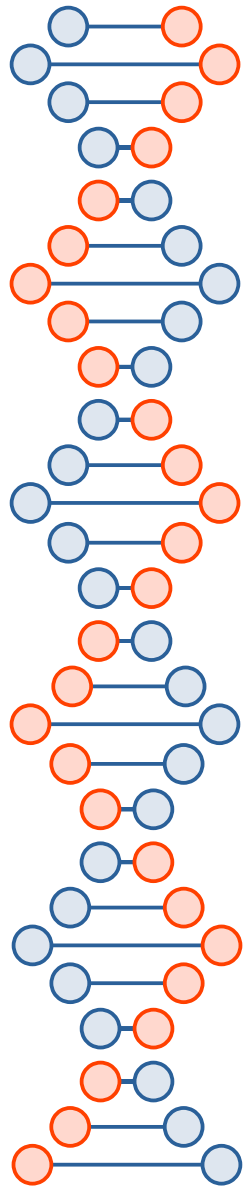
Étude des colonnes 'Indicator Name' et 'Indicator Code'

3 665 Indicateurs qui:

- Viennent de différentes bases de données (plus de 14 sources)
- Ne s'appliquent pas tous à tous les pays (il y a des indices 'spécifiques' et 'généraux')
- Compte parmi eux des indices de projections (vision future)

	Indicator Name	Indicator Code
0	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2
1	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F
2	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI
3	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M
...
3661	Youth literacy rate, population 15-24 years, b...	SE.ADT.1524.LT.ZS
3662	Youth literacy rate, population 15-24 years, f...	SE.ADT.1524.LT.FE.ZS
3663	Youth literacy rate, population 15-24 years, g...	SE.ADT.1524.LT.FM.ZS
3664	Youth literacy rate, population 15-24 years, m...	SE.ADT.1524.LT.MA.ZS

On choisira parmi ces indicateurs nos VARIABLES



Choix des Variables et Construction des jeux de données

Identification des variables de Projection

Parmi elles 2 semblent les plus intéressantes:

- PRJ.ATT.ALL.3.MF → % total ayant atteint Upper Secondary comme plus haut niveau
- PRJ.ATT.ALL.4.MF → % total ayant atteint Post Secondary ou Tertiary comme plus haut niveau

	Indicator Name	Indicator Code
3349	Wittgenstein Projection: Mean years of schooling. Age 0-19. Female	PRJ.MYS.0T19.FE
3350	Wittgenstein Projection: Mean years of schooling. Age 0-19. Male	PRJ.MYS.0T19.MA
3351	Wittgenstein Projection: Mean years of schooling. Age 0-19. Total	PRJ.MYS.0T19.MF
3352	Wittgenstein Projection: Mean years of schooling. Age 15+. Female	PRJ.MYS.15UP.FE
...
3653	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Primary. Total	PRJ.POP.ALL.1.MF
3654	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Female	PRJ.POP.ALL.3.FE
3655	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Male	PRJ.POP.ALL.3.MA
3656	Wittgenstein Projection: Population in thousands by highest level of educational attainment. Upper Secondary. Total	PRJ.POP.ALL.3.MF



Les projection ne concernent que des % de population ayant atteint certains niveaux scolaires et les temps moyens d'études.

Étude des deux variables de projection choisies

- Les variables de projection sont renseignées à partir de 2020.
- De 2020 a 2100 → des données pour estimer la projection dans le temps.

Années	Nombre de lignes renseignées	
	% Upper Secondary	% Post Secondary/Tertiary
1970-2009	0	0
2010	167	167
2011	0	0
2012	0	0
2013	0	0
2014	0	0
2015	167	167
2016	0	0
2017	0	0
2020-2100	167	167



Donc...

On gardera de 2020 à 2100 pour pouvoir faire le suivi dans le temps, c'est à dire la projection.

On garde que les colonnes de 2000 jusqu'à 2015 pour construire le modèle car il nous faut au moins 5 ans d'historique, et on a vu que cette période est celle avec le moins de données manquantes..

Jeu de Données de Projection

- 2 Variables de projection
- De 2020 à 2100
- 20 colonnes, 434 lignes (2 fois 217 pays. On enlève les 'Country Name' et 'Country Code' 'généraux', qui ont que des NaN).

Country Name	Country Code	Indicator Name	Indicator Code	2020	2025	2030	2035	2040	2045	...	2055	2060	2065	2070	2075	2080	2085	2090	2095	2100
Arab World	ARB	Wittgenstein Projection: Percentage of the total population by highest level of educational attainment. Post Secondary. Total	PRJ.ATT.ALL.4.MF	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...
Zimbabwe	ZWE	Wittgenstein Projection: Percentage of the total population by highest level of educational attainment. Upper Secondary. Total	PRJ.ATT.ALL.3.MF	0.32	0.37	0.41	0.44	0.47	0.5	...	0.54	0.56	0.57	0.58	0.59	0.59	0.59	0.59	0.59	0.58

Identification des variables Historiques

- En filtrant les NaN et les dupliqués pour 2000-2015 on a 561 indices parmi lesquels on peut choisir nos variables (un plus vaste choix qu'en incluant 2016 qui limite à 153 indices)

	Indicator Name	Indicator Code
2483	Population, ages 15-64 (% of total)	SP.POP.1564.TO.ZS
4907	GDP at market prices (constant 2005 US\$)	NY.GDP.MKTP.KD
4908	GDP at market prices (current US\$)	NY.GDP.MKTP.CD
4909	GDP per capita (constant 2005 US\$)	NY.GDP.PCAP.KD
...
790917	School life expectancy, tertiary, male (years)	UIS.SLE.56.M
790951	Teachers in tertiary education ISCED 5 programmes, both sexes (number)	UIS.T.5.B
790952	Teachers in tertiary education ISCED 5 programmes, female (number)	UIS.T.5.B.F
791063	Total inbound internationally mobile students, both sexes (number)	UIS.MS.56.T

561 rows × 2 columns



Choix des variables

- % de public niveau lycée/ population totale
 - Gross enrolment ratio, secondary
 - Lower secondary completion rate
- % de public niveau université/ population totale
 - Gross enrolment ratio, tertiary
- Connexion Internet
 - Internet users
- Pouvoir d'achat (payer inscription et ordinateur)
 - GNI per capita, Atlas method
 - Labor force with advanced education
 - Labor force with intermediate education
 - Expenditure on education as % of total government expenditure.
 - Expenditure on tertiary as % of government expenditure on education

Identification des variables Historiques

	Indicator Name	Indicator Code
6	1224 Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS
7	1238 Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS
5	1251 GNI per capita, Atlas method (current US\$)	NY.GNP.PCAP.CD
1	1335 Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR
3	1339 Gross enrolment ratio, tertiary, both sexes (%)	SE.TER.ENRR
4	1375 Internet users (per 100 people)	IT.NET.USER.P2
8	1376 Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS
9	1382 Labor force with intermediate education (% of total)	SL.TLF.INTM.ZS
2	1518 Lower secondary completion rate, both sexes (%)	SE.SEC.CMPT.LO.ZS



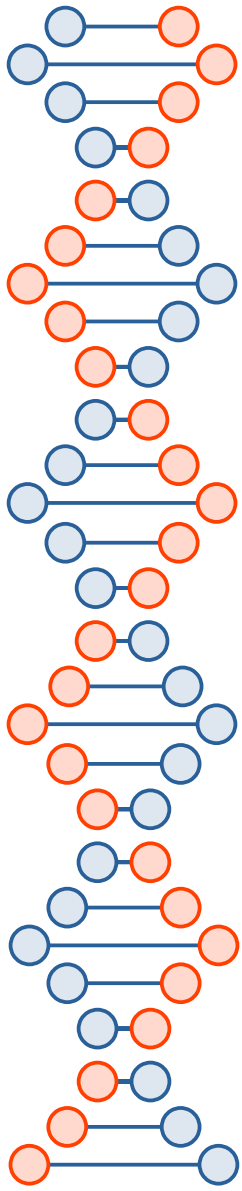
En ordre d'importance les variables sont:
GSec, CLSec, GTer, Net, GNI, ExpEd, ExpTer, LAdv, LInt

Jeu de données Historiques

- 9 Variables de projection
- De 2000 à 2015
- 2178 lignes, 20 colonnes

	Country Name	Country Code	Indicator Name	Indicator Code	2000	...	2011	2012	2013	2014	2015
1224	Arab World	ARB	Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
1238	Arab World	ARB	Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
1251	Arab World	ARB	GNI per capita, Atlas method (current US\$)	NY.GNP.PCAP.CD	2388.342883	...	6305.644064	7196.532736	NaN	NaN	NaN
...
884641	Zimbabwe	ZWE	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
884647	Zimbabwe	ZWE	Labor force with intermediate education (% of total)	SL.TLF.INTM.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
884783	Zimbabwe	ZWE	Lower secondary completion rate, both sexes (%)	SE.SEC.CMPT.LO.ZS	NaN	...	NaN	64.476013	65.527359	NaN	NaN

2178 rows × 20 columns



Analyse du jeu de données Historique (2000-2015)

Par pays

(ie. sans les 'Country Codes' qui correspondent à des lignes 'générales')

	Country Name	Country Code	Indicator Name	Indicator Code	2000	...	2011	2012	2013	2014	2015
92849	Afghanistan	AFG	Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS	NaN	...	16.048429	10.356800	14.102800	14.465930	12.509000
92863	Afghanistan	AFG	Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	NaN	...	8.986210	12.741710	11.756830	12.411280	15.953790
92876	Afghanistan	AFG	GNI per capita, Atlas method (current US\$)	NY.GNP.PCAP.CD	NaN	...	560.000000	670.000000	670.000000	630.000000	590.000000
92960	Afghanistan	AFG	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	NaN	...	54.616180	56.677341	56.688660	55.656158	55.644409
...
884640	Zimbabwe	ZWE	Internet users (per 100 people)	IT.NET.USER.P2	0.401434	...	8.400000	12.000000	15.500000	16.364740	22.742818
884641	Zimbabwe	ZWE	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
884647	Zimbabwe	ZWE	Labor force with intermediate education (% of total)	SL.TLF.INTM.ZS	NaN	...	NaN	NaN	NaN	NaN	NaN
884783	Zimbabwe	ZWE	Lower secondary completion rate, both sexes (%)	SE.SEC.CMPT.LO.ZS	NaN	...	NaN	64.476013	65.527359	NaN	NaN

1953 rows × 20 columns

pour nos variables il y a des lignes qui ont des NaN dans toutes les années, et d'autres qui on des NaN juste pour certaines années.

Statistiques pour nos variables par Pays

(moyennes, écarts types, médianes)

	Country Name	Country Code	Indicator Name	Indicator Code	...	2015	Mean	Std	Median
92849	Afghanistan	AFG	Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS	...	12.509000	14.09	2.21	14.10
92863	Afghanistan	AFG	Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	...	15.953790	11.81	2.38	11.78
92876	Afghanistan	AFG	GNI per capita, Atlas method (current US\$)	NY.GNP.PCAP.CD	...	590.000000	458.33	162.06	459.16
92960	Afghanistan	AFG	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	...	55.644409	38.88	16.83	39.55
...
884640	Zimbabwe	ZWE	Internet users (per 100 people)	IT.NET.USER.P2	...	22.742818	6.43	6.51	3.75
884641	Zimbabwe	ZWE	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	...	NaN	NaN	NaN	NaN
884647	Zimbabwe	ZWE	Labor force with intermediate education (% of total)	SL.TLF.INTM.ZS	...	NaN	NaN	NaN	NaN
884783	Zimbabwe	ZWE	Lower secondary completion rate, both sexes (%)	SE.SEC.CMPT.LO.ZS	...	NaN	65.00	0.53	64.74

1953 rows × 23 columns

Établissement du Score à points

Système de points:

- 1er quartile ->0
- 2ème quartile ->1
- 3ème quartile ->2
- 4ème quartile ->3

Score maximum $3 \cdot 9 = 27$

	Country Name	Country Code	Score_GSec	Score_CLSec	Score_GTer	Score_Net	Score_GNI	Score_ExpEd	Score_ExpTer	Score_LAdv	Score_LInt	Score_Total
0	Afghanistan	AFG	0	0	0	0	0	1	0	0	0	1
1	Albania	ALB	2	2	2	2	1	0	1	1	1	12
2	Algeria	DZA	1	1	1	1	1	0	2	0	0	7
3	American Samoa	ASM	0	0	0	0	0	0	0	0	0	0
...
13	West Bank and Gaza	PSE	2	1	2	2	1	0	0	1	0	9
14	Yemen, Rep.	YEM	0	0	1	1	0	3	0	0	0	5
15	Zambia	ZMB	0	1	0	0	0	0	1	0	0	2
16	Zimbabwe	ZWE	0	1	0	0	0	3	1	0	0	5

17 rows × 12 columns

Définition du meilleur Score

La moyenne étant à 10.3 on peut dire que au de là de **15** on considère un Score suffisamment bon. →

count	217.000000
mean	10.290323
std	6.853193
min	0.000000
25%	4.000000
50%	10.000000
75%	15.000000
max	26.000000

Les 20 Pays au meilleur Score

	Country Name	Country Code	Score_GSec	Score_CLSec	Score_GTer	Score_Net	Score_GNI	Score_ExpEd	Score_ExpTer	Score_LAdv	Score_LInt	Score_Total
53	Denmark	DNK	3	3	3	3	3	2	3	3	3	26
146	Norway	NOR	3	3	3	3	3	2	3	3	3	26
87	Iceland	ISL	3	3	3	3	3	2	2	3	3	25
92	Ireland	IRL	3	3	3	3	3	1	3	3	3	25
...
178	Spain	ESP	3	2	3	3	3	0	2	3	3	22
10	Australia	AUS	3	0	3	3	3	1	2	3	3	21
52	Czech Republic	CZE	3	3	3	3	3	0	2	2	2	21
67	France	FRA	3	3	3	3	3	0	2	2	2	21

20 rows × 12 columns

Les 50 Pays au meilleur Score

	Country Name	Country Code	Score_GSec	Score_CLSec	Score_GTer	Score_Net	Score_GNI	Score_ExpEd	Score_ExpTer	Score_LAdv	Score_LInt	Score_To
53	Denmark	DNK	3	3	3	3	3	2	3	3	3	
146	Norway	NOR	3	3	3	3	3	2	3	3	3	
87	Iceland	ISL	3	3	3	3	3	2	2	3	3	
92	Ireland	IRL	3	3	3	3	3	1	3	3	3	
...	
115	Luxembourg	LUX	3	3	1	3	3	0	0	2	1	
192	Thailand	THA	2	1	2	1	1	3	2	2	2	
207	Uruguay	URY	3	1	2	2	2	0	2	1	3	
9	Aruba	ABW	3	3	2	3	0	3	1	0	0	

50 rows × 12 columns

Si on prend les 50 pays avec meilleur score on retrouve des pays de:

- Europe & Central Asia (ECS) and European Union (EUU),
- North America (NAC),
- Latin America & Caribbean (LCN) and
- East Asia & Pacific(EAS)

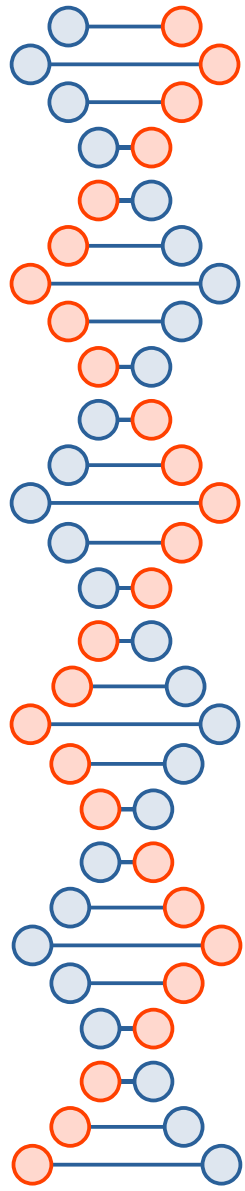
Dans cet ordre, les zones géographiques dont les statistiques dégagent le plus de potentiel (à voir par la suite). On gardera donc le 50 pays avec le meilleur score...

Par zone géographique

(ie. on garde que les lignes 'générales')

	Country Name	Country Code	Indicator Name	Indicator Code	...	2015	Mean	Std	Median
4889	East Asia & Pacific	EAS	Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS	...	NaN	NaN	NaN	NaN
4903	East Asia & Pacific	EAS	Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	...	NaN	NaN	NaN	NaN
4916	East Asia & Pacific	EAS	GNI per capita, Atlas method (current US\$)	NY.GNP.PCAP.CD	...	9798.860171	6427.944305	2206.474953	5902.280984
...
78341	Sub-Saharan Africa	SSF	Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	...	NaN	NaN	NaN	NaN
78347	Sub-Saharan Africa	SSF	Labor force with intermediate education (% of total)	SL.TLF.INTM.ZS	...	NaN	NaN	NaN	NaN
78483	Sub-Saharan Africa	SSF	Lower secondary completion rate, both sexes (%)	SE.SEC.CMPT.LO.ZS	...	NaN	35.200332	4.588932	34.741776

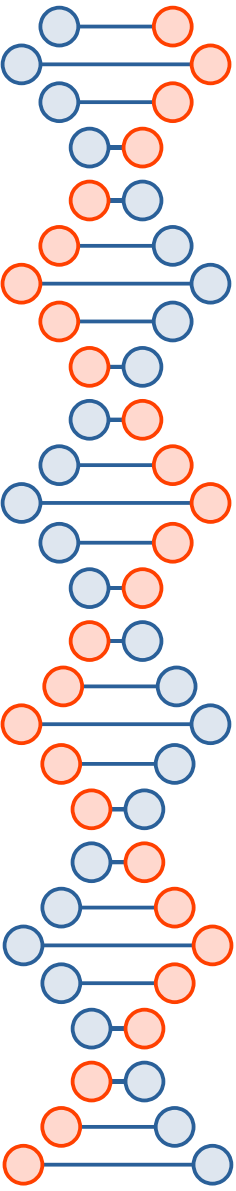
72 rows × 23 columns



Column	Nb. of values
Country Name	72
Country Code	72
Indicator Name	72
Indicator Code	72
2000	41
2001	41
2002	41
2003	41
2004	41
2005	41
2006	41
2007	41
2008	41
2009	45
2010	43
2011	45
2012	45
2013	45
2014	41
2015	18
Mean	45
Std	45
Median	45

Qualité du jeu de données par zone géographique

- On a 8 Zones Géographiques et 9 variables soit $8 \cdot 9 = 72$ lignes.
- Pas toutes les variables sont renseignées et surtout en 2015 on a beaucoup plus de non renseignés. Pour étudier les statistiques de nos variables choisies, lorsque c'est possible, par zones géographiques on va donc enlever 2015.

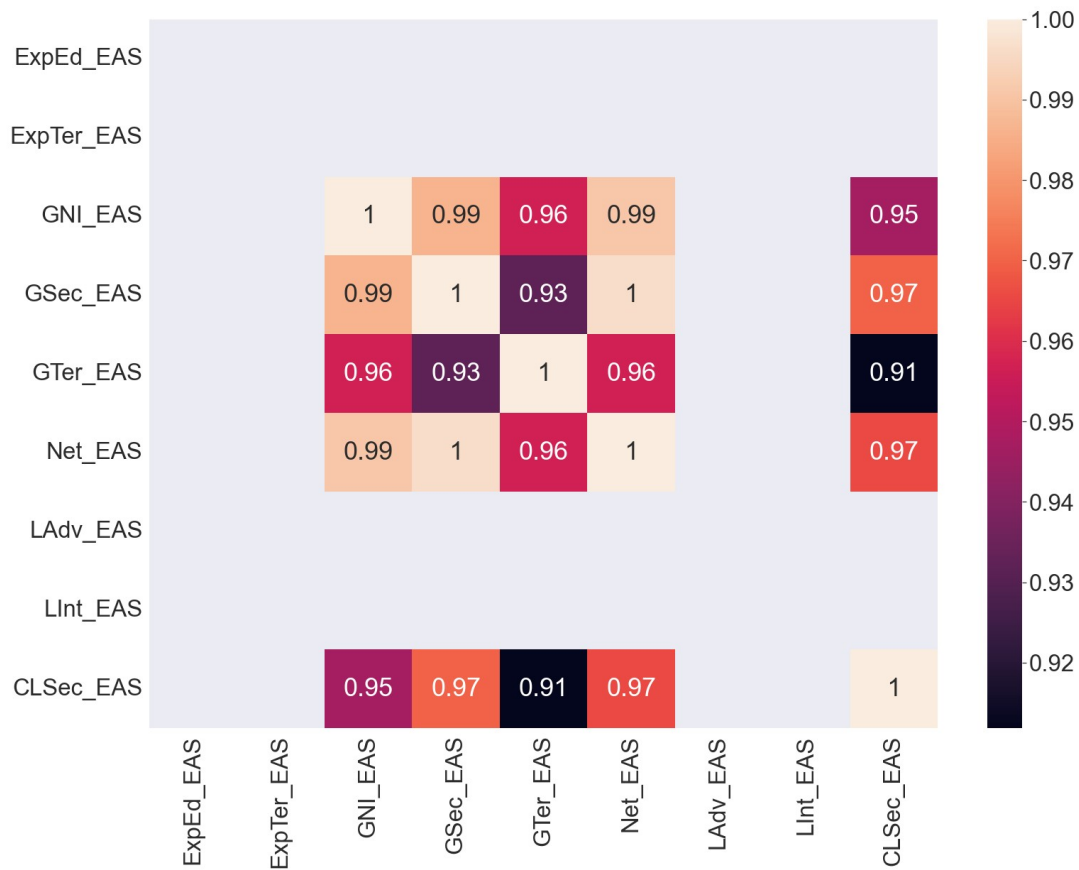


Graphiques des corrélations

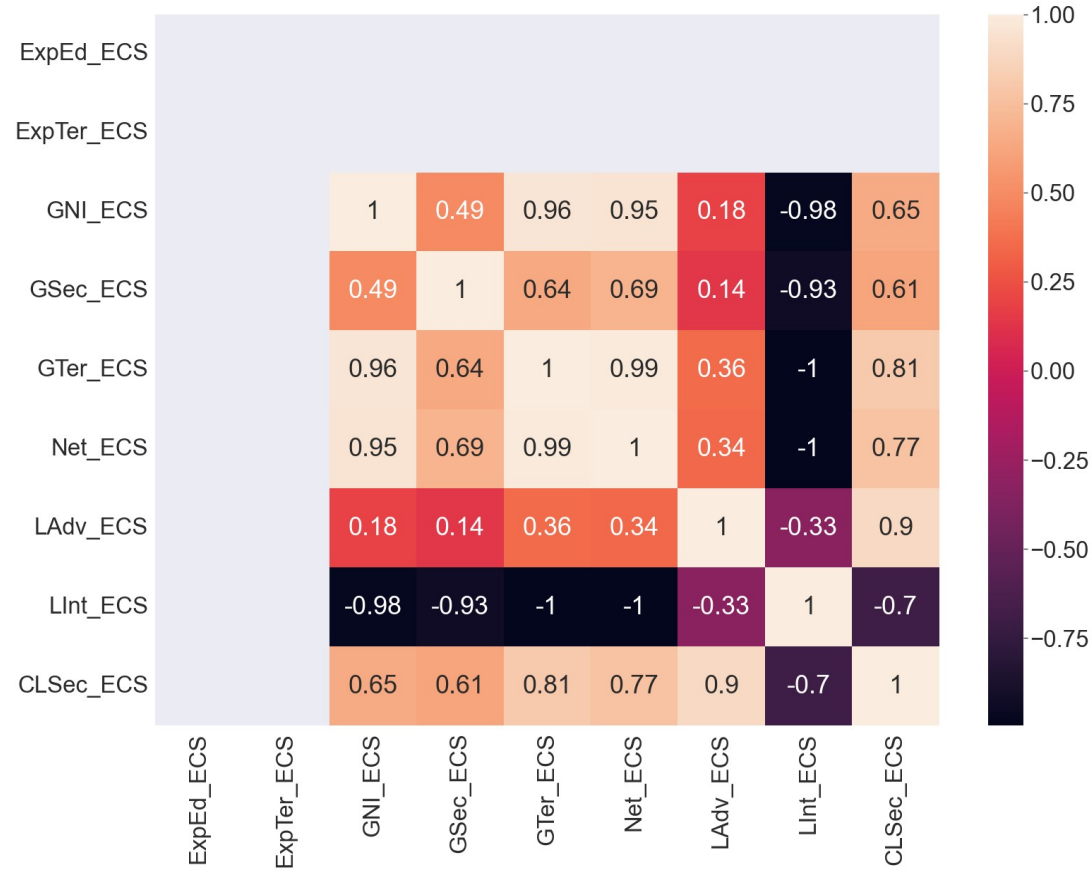
(toutes les zones géographique ensemble)

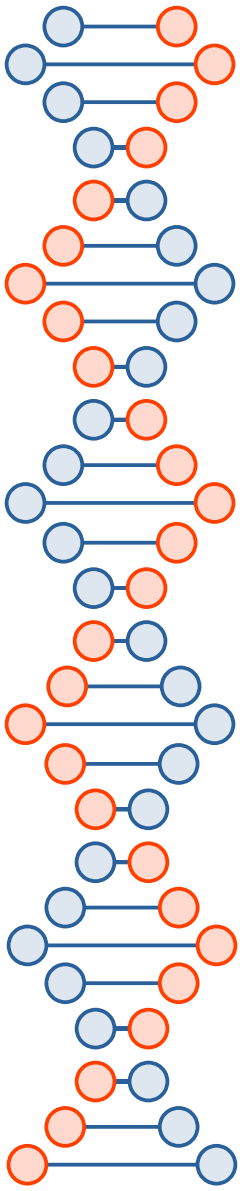


East Asia & Pacific

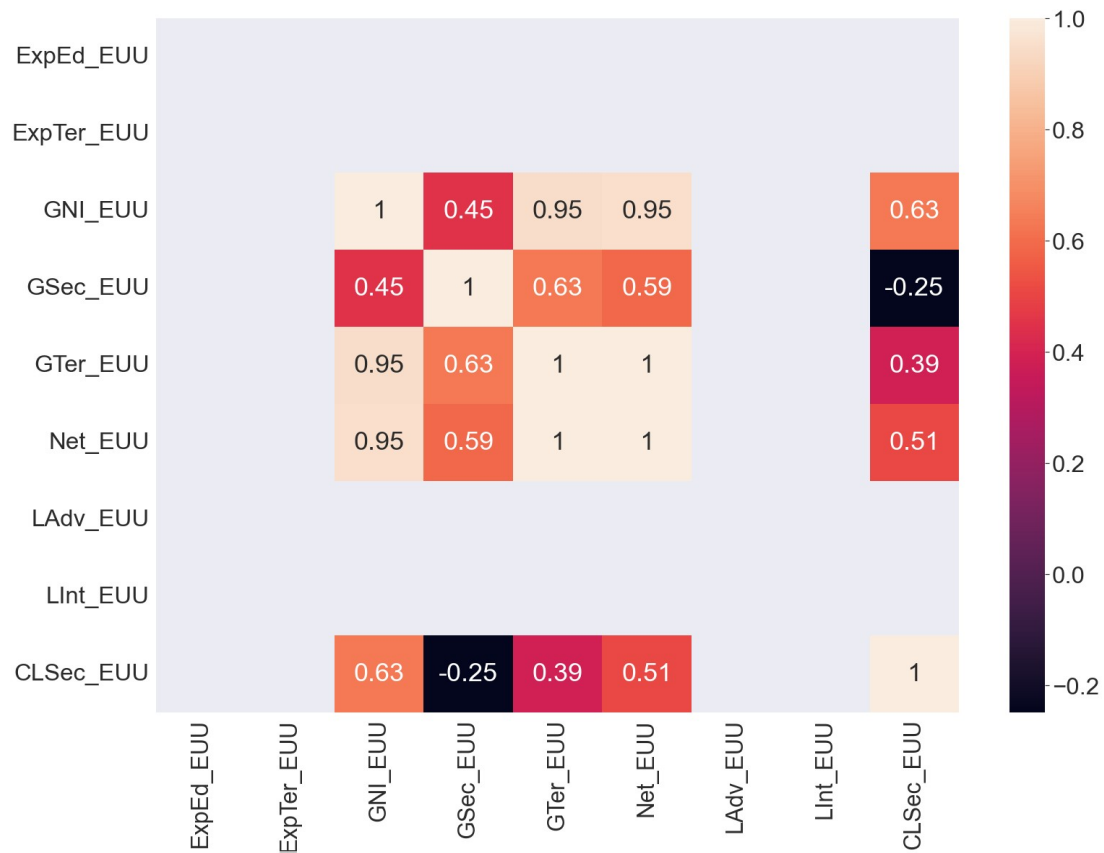


Europe & Central Asia

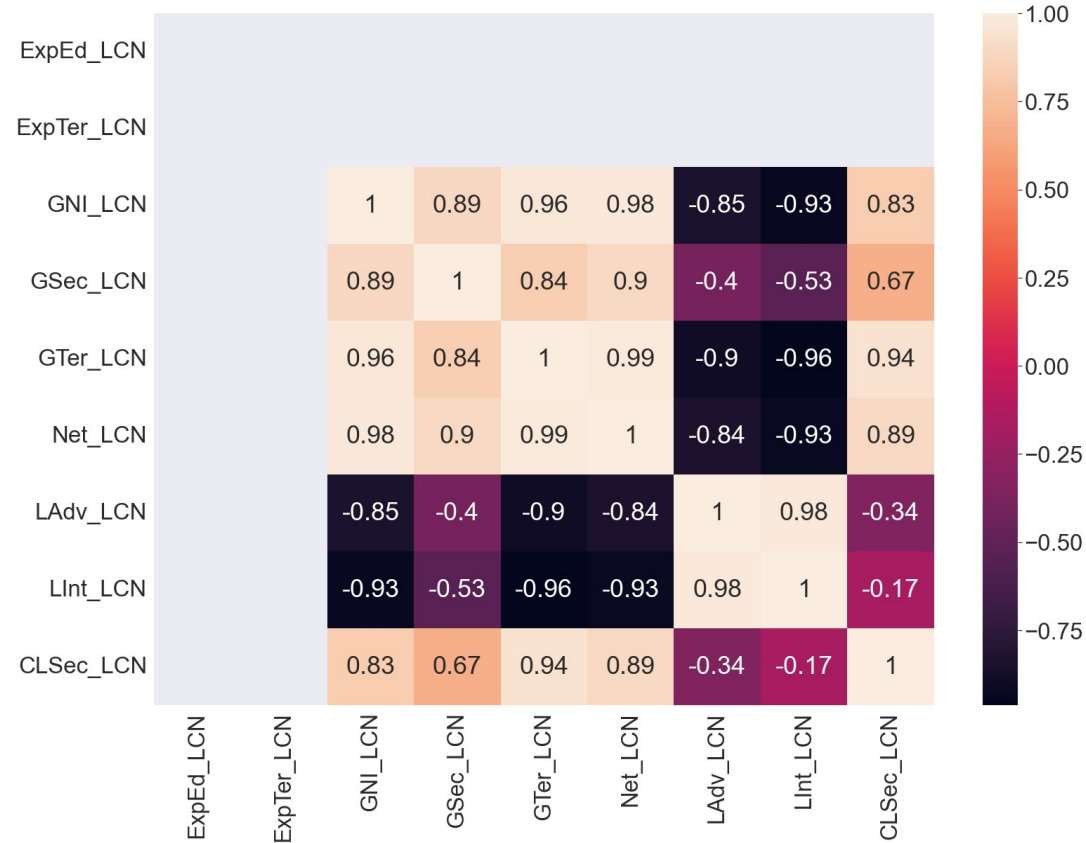


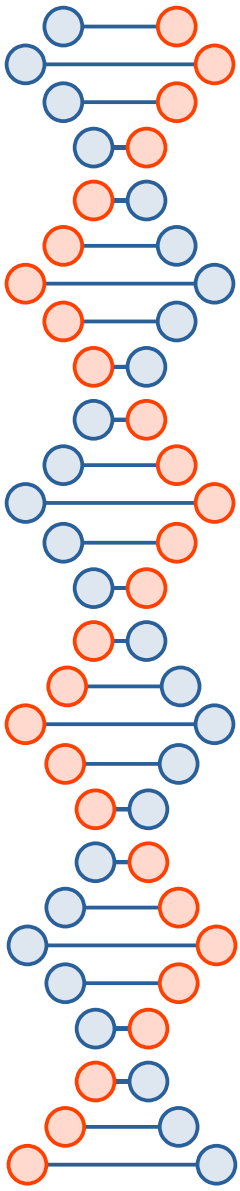


European Union



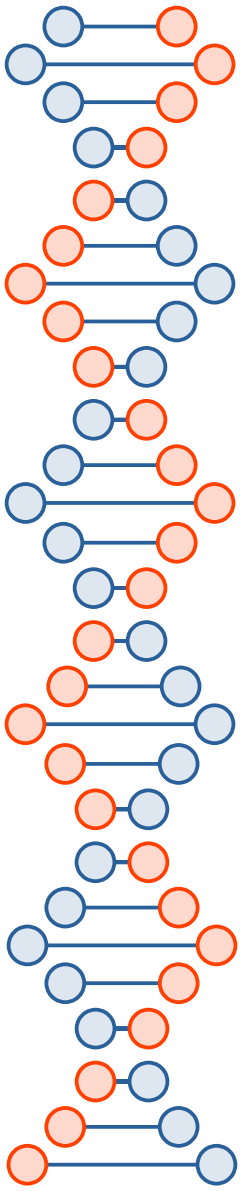
Latin America & Caribbean



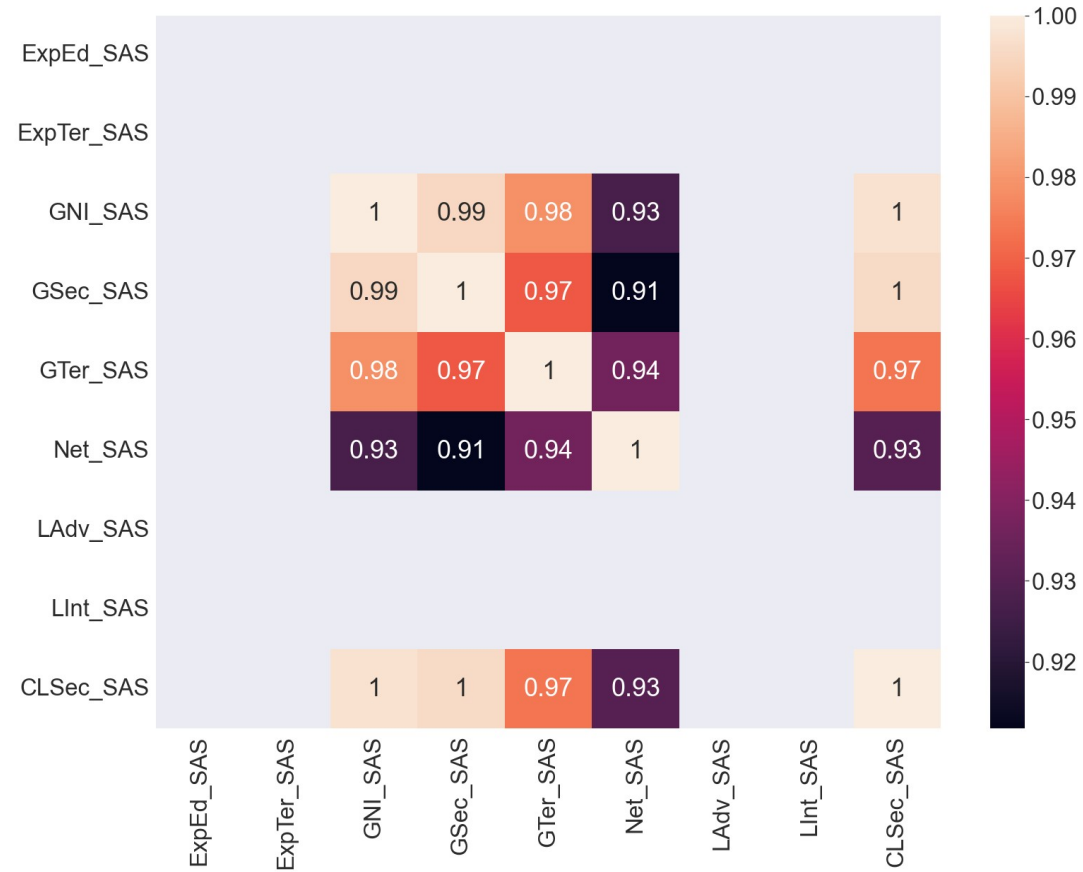


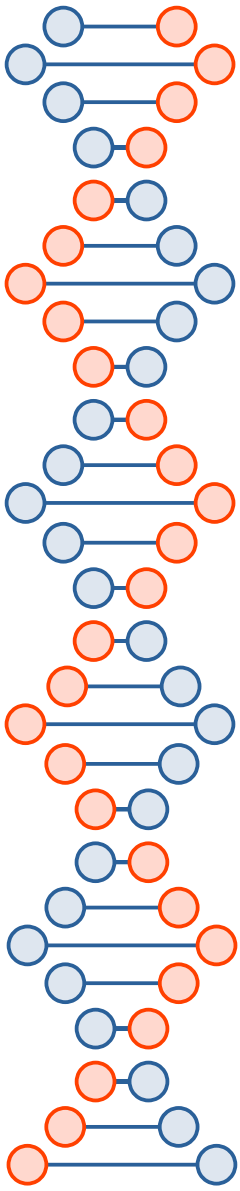
North America



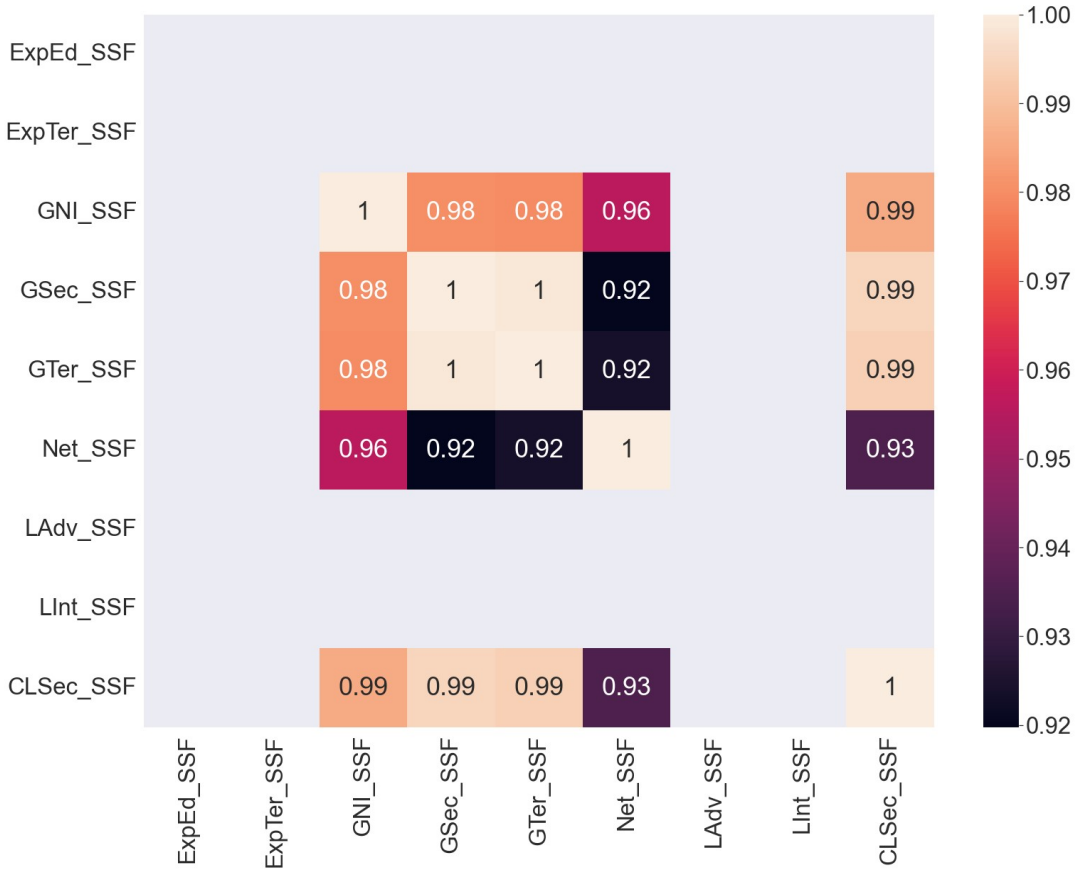


South Asia





Sub-Saharan Africa



Résultats des corrélations

- Dans les pays « développés » (ie. 1^{er} monde) les études, l'accès à internet et le budget ne sont pas corrélés : On peut être pauvres et avoir un haut niveau d'études.
C'est m'inverse dans les pays pauvres et en voie de développement.
- Comme on travaille avec un Score à points on peut se permettre de travailler avec des variable très corrélées entre elles.
(Score à dire d'expert)

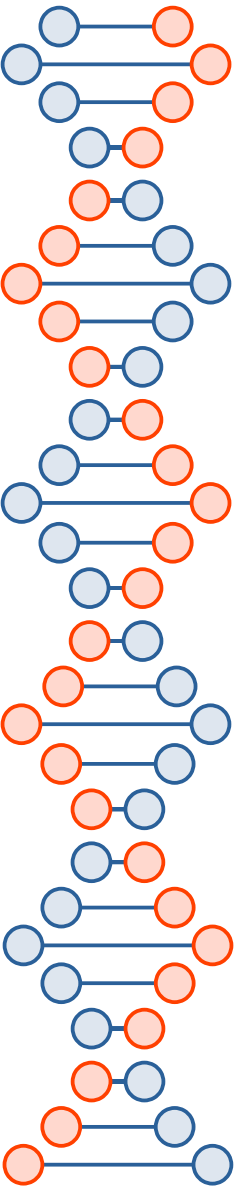


Histogrammes comparés

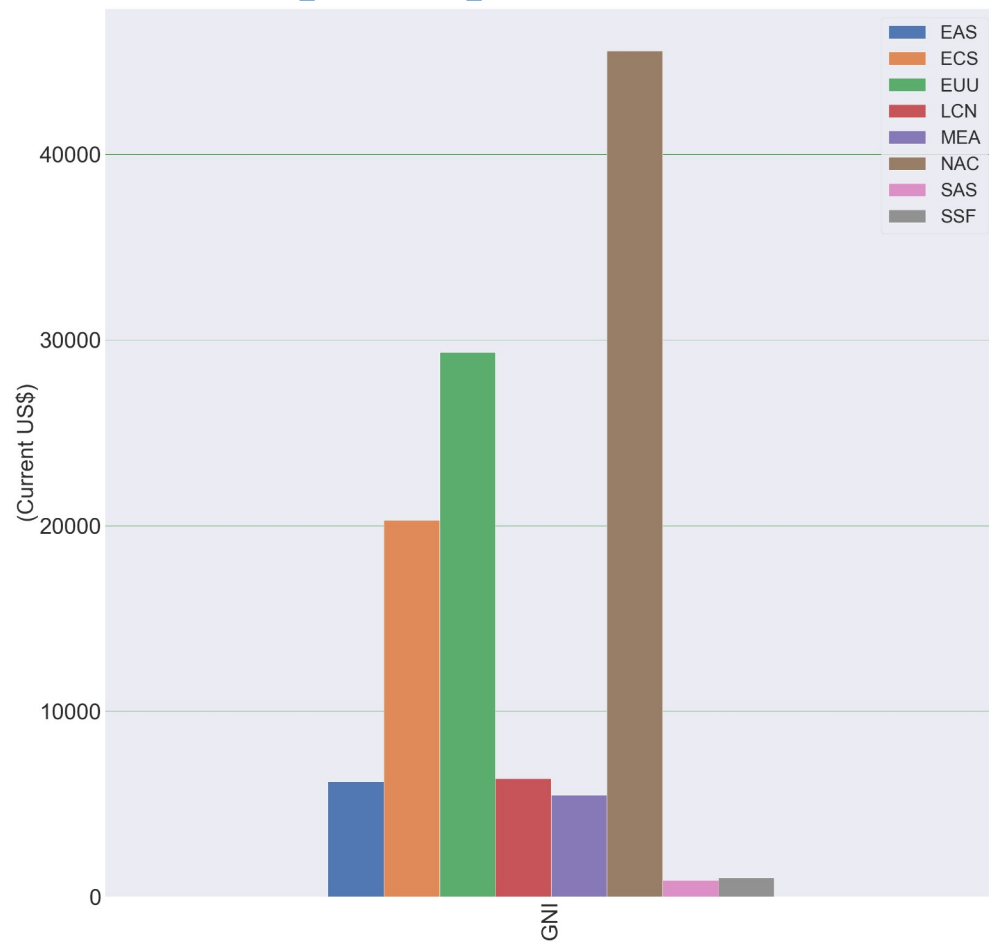
(par Zones Géographiques des Moyennes des 4 principaux indices: GNI, GSec, GTer, Net et CLSec)

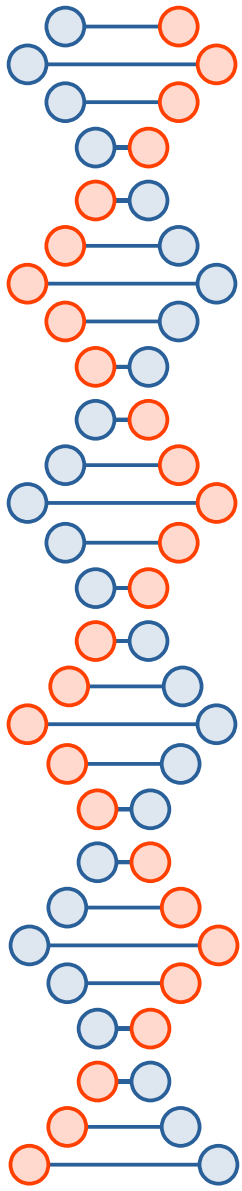


ExpEd, ExpTer, ne sont pas renseignés pour les Zones Géographique et LAdv, LInt, sont renseignés uniquement pour ECS, LCN et NAC.

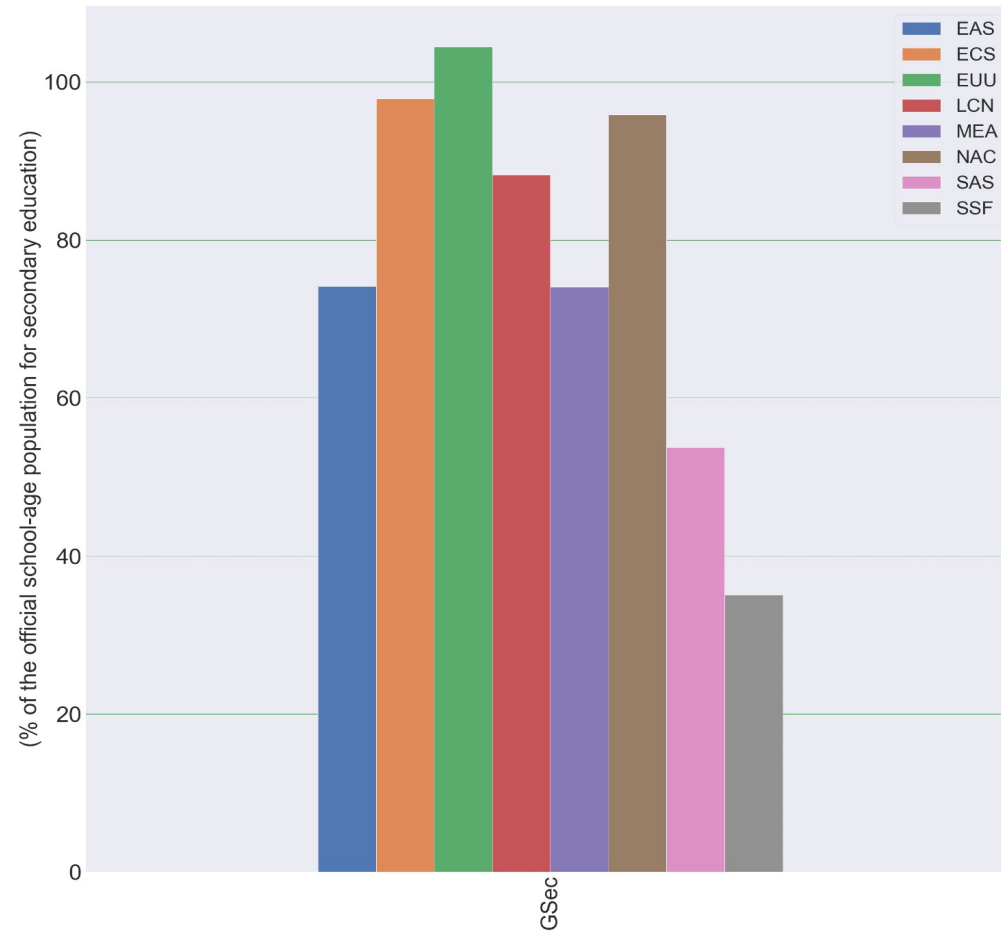


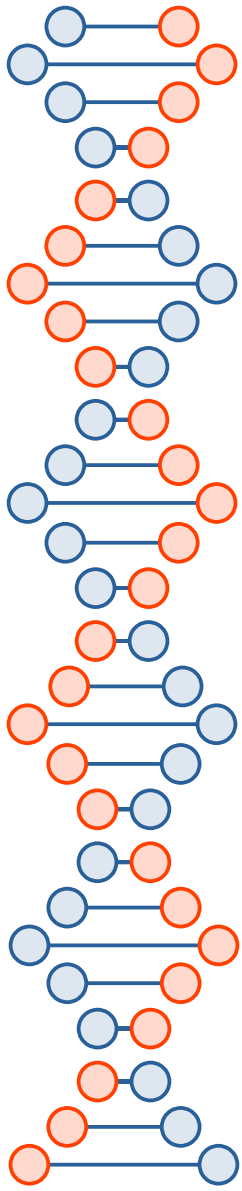
GNI per capita, Atlas method



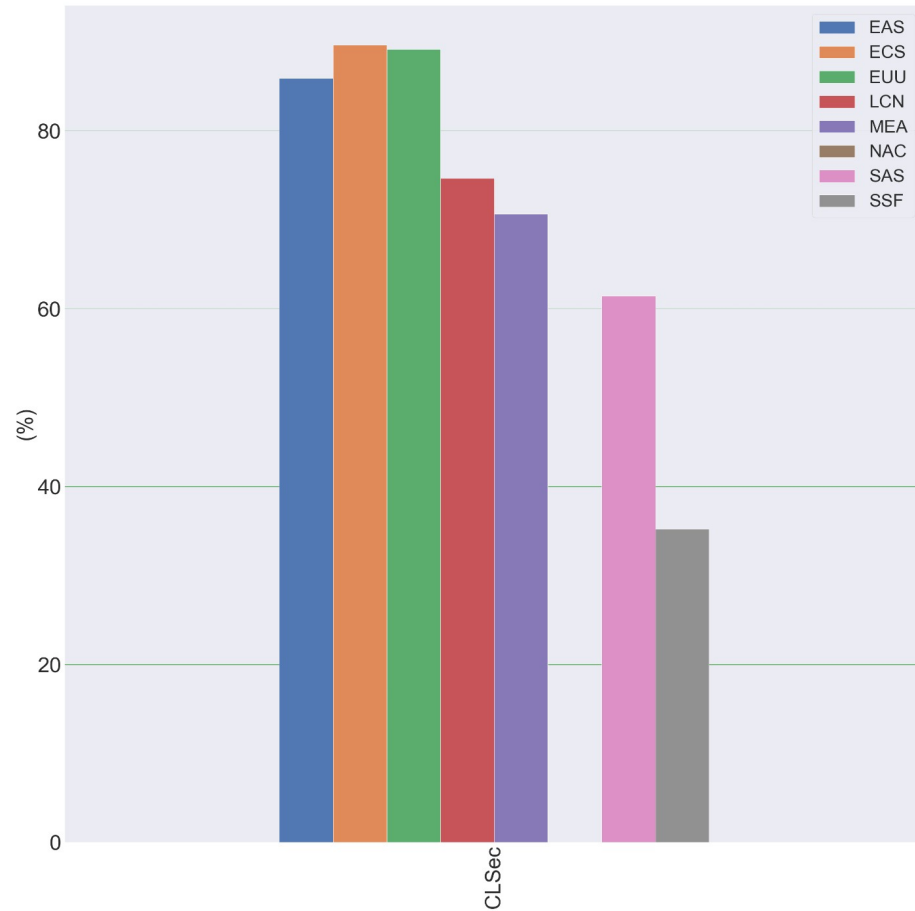


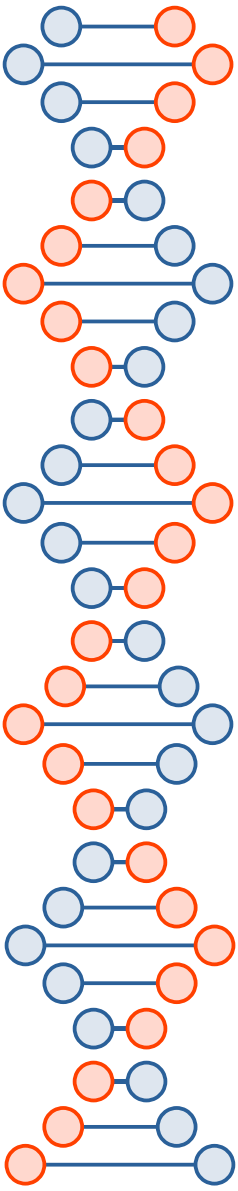
Gross enrolment ratio, secondary



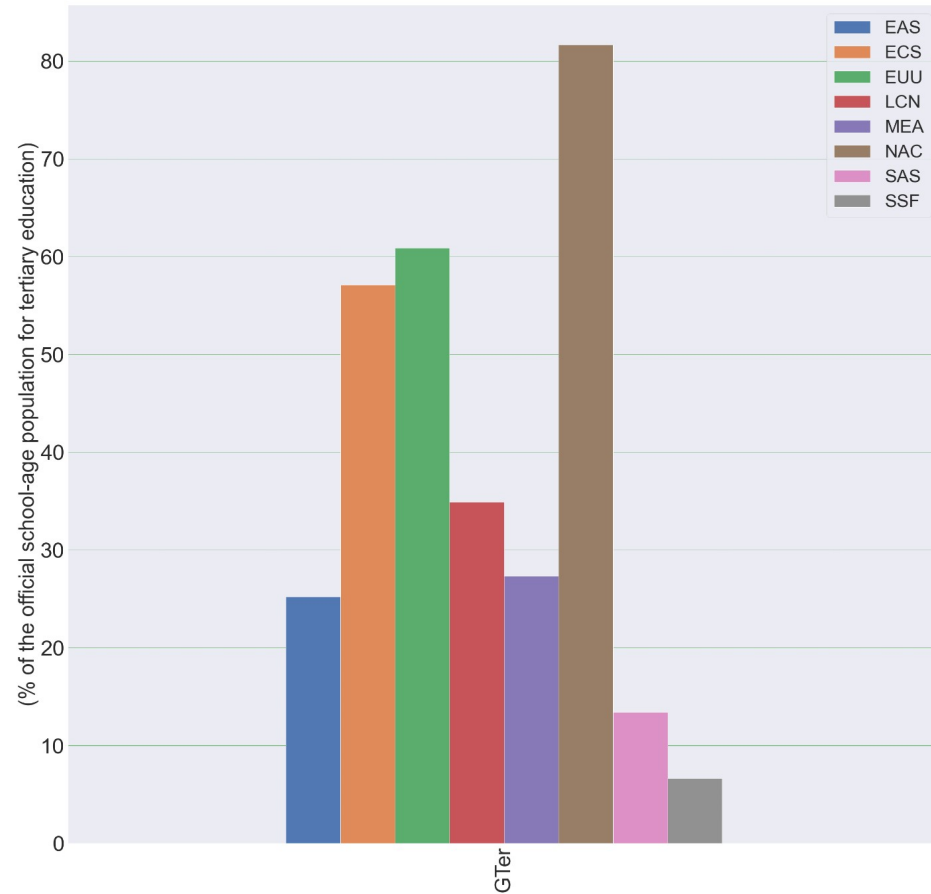


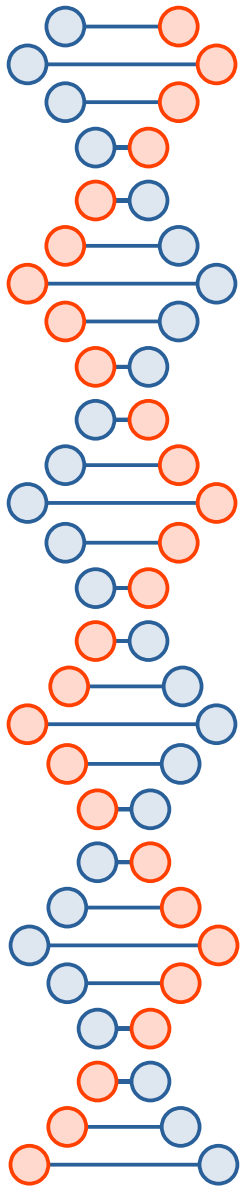
Lower secondary completion rate



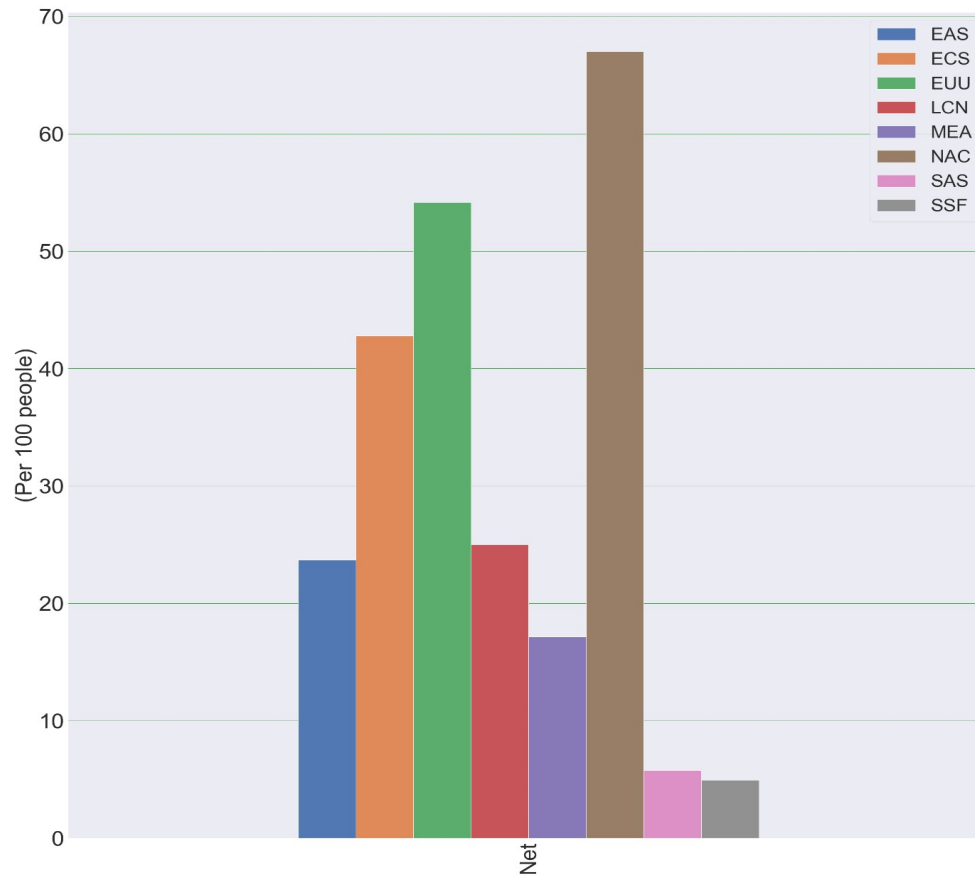


Gross enrolment ratio, tertiary





Internet users

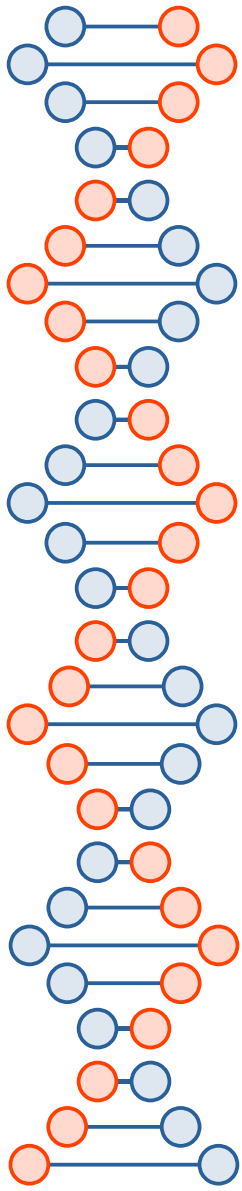


Résultats des histogrammes

- L'Europe (en orange et vert) et l'Amérique du Nord présentent les meilleurs comportements.
- Ne pouvant pas nous restreindre à ces zones du premier monde (il faut s'ouvrir aux pays en développement à fort potentiel) les zones de l'Amérique latine et les Caraïbes et de l'Asie de l'Est et du Pacifique représentent un terrain intéressant.
(Ceci s'est vérifié en regardant les 50 pays à meilleur Score et le sera aussi avec les projection qui sont à suivre).
- L'Afrique Subsaharienne, l'Asie du Sud et le Moyen Orient et l'Afrique du Nord, ne sont pas des zones intéressantes, car elle présentent les handicapés économiques et de réseau internet.

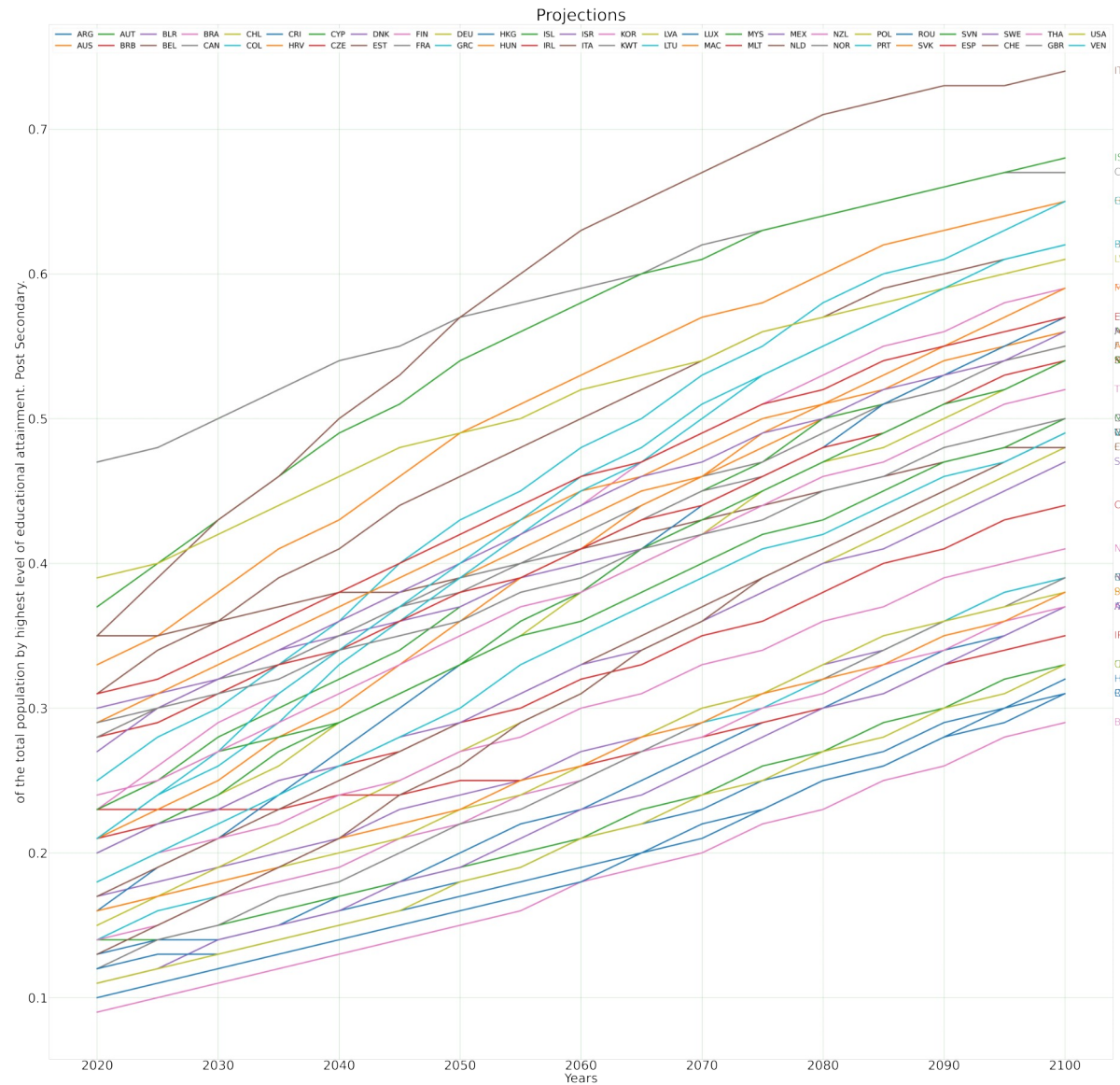


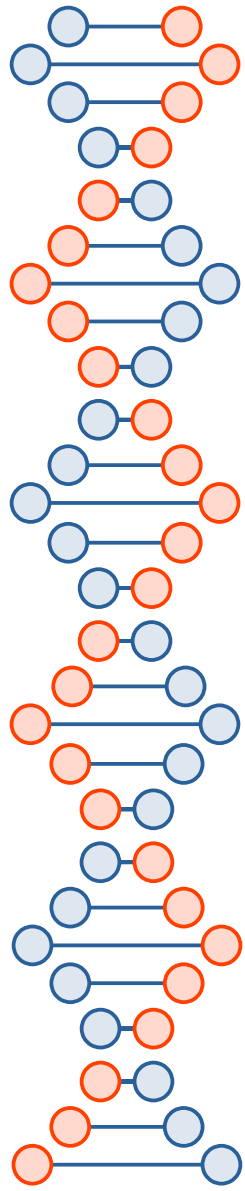
Projection futures les 50 pays au meilleur Score



% ayant pour
plus haut
niveau
d'éducation
atteint Post
Secondary

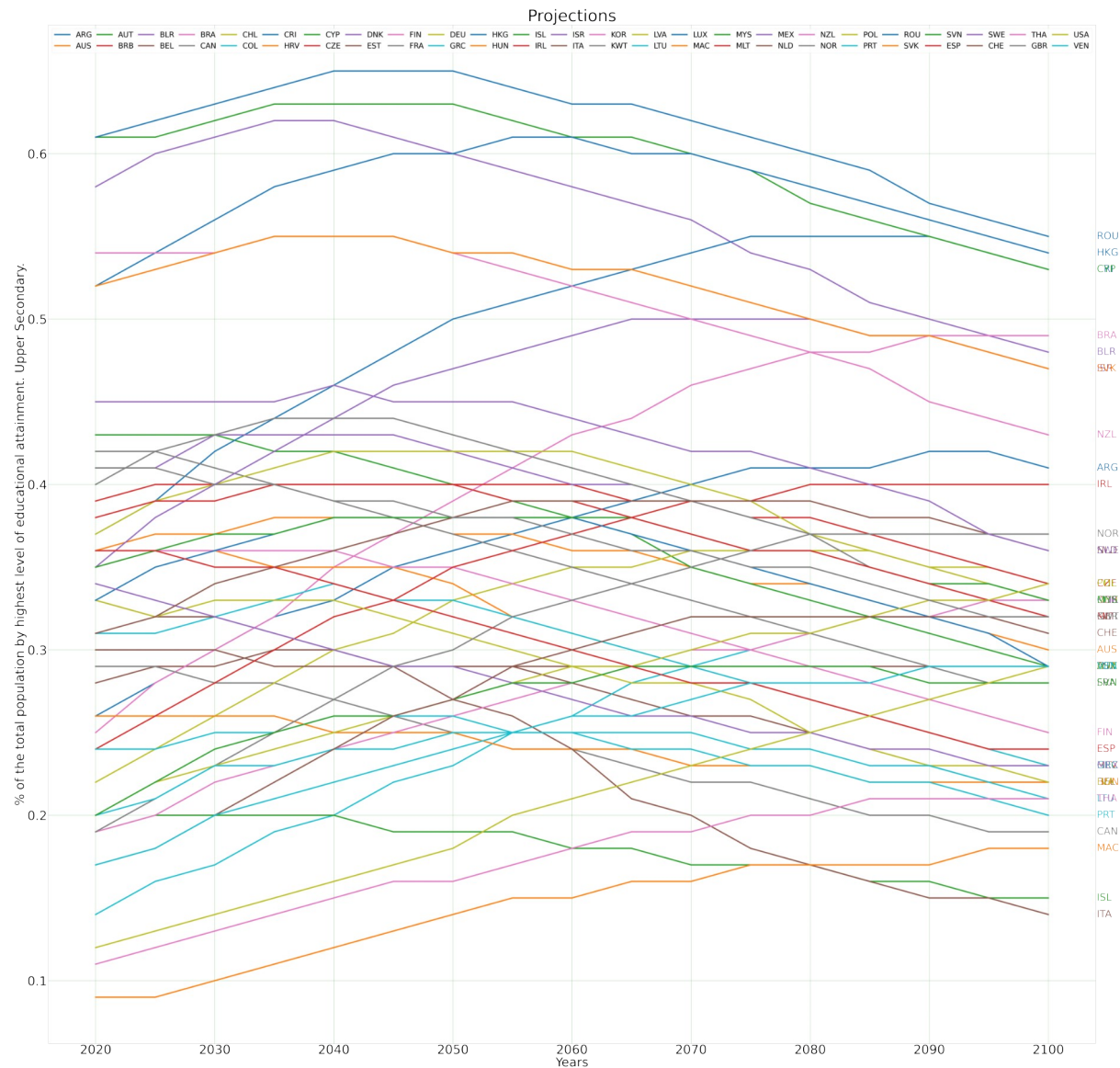
Tous les pays ont une
projection croissante.

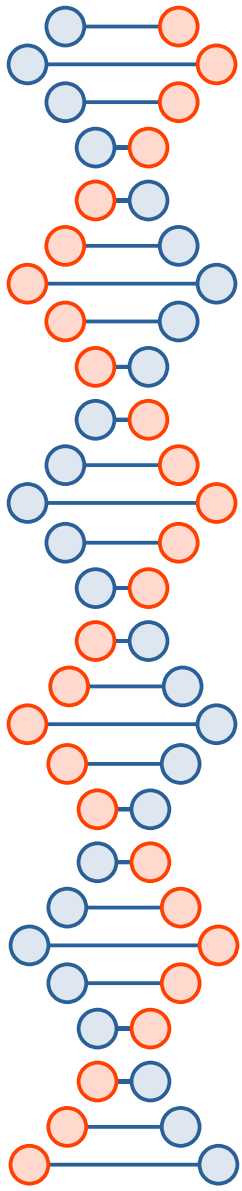




% ayant pour plus haut niveau d'éducation atteint Upper Secondary

Uniquement: ARG, BRA, CHL, COL, CRI, HKG, IRL, ISR, KOR, MAC, NLD, NOR, POL, SVN, CHE, THA, USA, VEN ont une projection croissante. Ce sont des pays de LCN, ECS, EAS et NAC; ce qui confirme le fait d'investir dans ces Zones Géographiques.





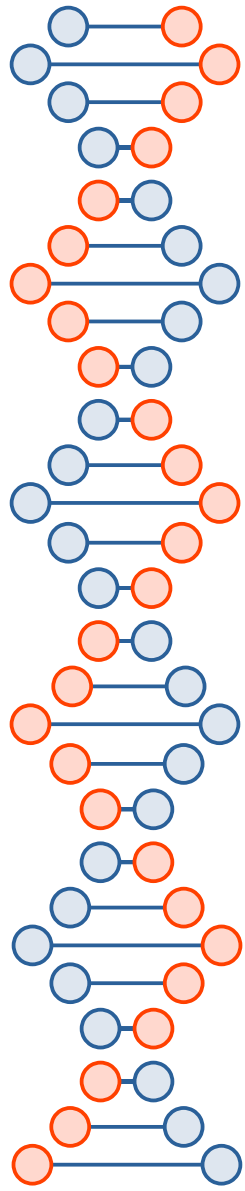
CONCLUSIONS



Dans quels pays l'entreprise doit-elle opérer en priorité ?

Argentine, Brésil, Chili, Colombie, Costa Rica, Hong Kong, Irlande, Israël, République de Corée, Macao, Pays-Bas, Norvège, Pologne, Slovaquie, Suisse, Thaïlande, États-Unis, Venezuela.

Ce sont les pays parmi les 50 meilleurs scores (et donc meilleur potentiel) et appartenant aux zones géographiques à comportement intéressant qui possèdent une projection croissante dans nos deux indices de projection.



MERCI