Analyse des données de systèmes éducatifs

Pré-analyse pour la start-up *academy*

Problématiques

– Les pays avec un fort potentiel de clients ?

– Quelle sera l'évolution de ce potentiel ?

– Quels sont les pays prioritaires ?

Plan

I) La base de données Edstats

La base de donnée contient 3665 indicateurs pour 241 pays.

— EdStatsCountry.csv

— EdStatsCountry-Series.csv

— EdStatsData.csv

— EdStatsFootNote.csv

— EdStatsSeries.csv

n_col	n_lin
31	241
4	613
69	886930
5	643638
20	3665

I) 1-EdStatsCountry.csv - country

Contient des informations générales sur chaque pays.

Ligne = Pays (241)

– Monnaie

– Groupes internationaux

– …

Clés primaires : Country Code, Short Name...

I) 2-EdStatsSeries.csv - series

Renseigne sur les indicateurs disponibles.

Ligne = Indicateur (3665)

– Identifiant

– Thème

– Définitions

– Périodicité

Clés primaires : Series Code, Indicator Name...

I) 3-EdStatsCountry-Series.csv

Donne des informations sur la source pour un indicateur et un pays.

Ligne = Indicateur pour un pays (613)

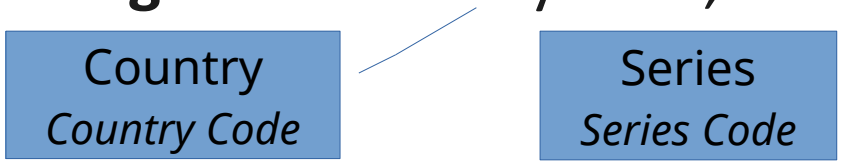
– Méthode

– Sélection des données

– …

Clé primaire: CountryCode & SeriesCode

Clés étrangères : CountryCode, SeriesCode



I) 5-EdStatsData.csv - data

Contient les données des indicateurs pour chaque pays.

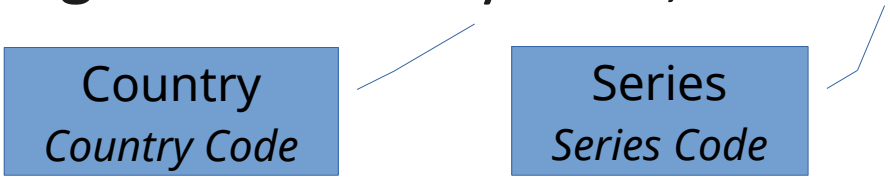
Ligne = Valeurs d'un indicateur pour un pays pour les années entre 1970 et 2100 (886930)

– Identifiants indicateurs, pays

– Valeurs indicateurs

Clé primaire : Country Code & Indicator Code, ...

Clés étrangères : Country Code, Indicator Code



I) 4-EdStatsFootNote.csv

Donne des informations sur la source pour un indicateur et un pays pour une année.

Ligne = Indicateur pour un pays pour une année (643638)

– Source

– …

Clé primaire : CountryCode & SeriesCode & Year

Clés étrangères : CountryCode, SeriesCode

II) Environnement de travail

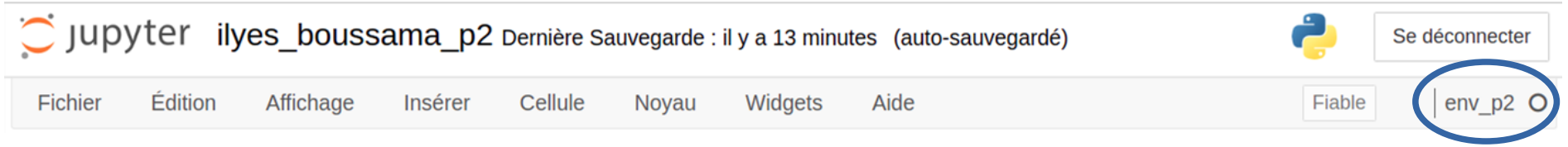
conda create --name env_p2

conda -n env_p2 install python

conda activate env_p2

python -m ipykernel install --user –name=env_p2

conda deactivate



III) Sélection des indicateurs

Les indicateurs doivent nous renseigner sur:

– La demande de formation

– Les possibilités matérielles/financières

Méthode :

– Filtrage des indicateurs par mots clés

– Sélection en fonction de la quantité moyenne de données disponible entre 2004 et 2014

III) Indicateurs choisis

— Gross enrolment ratio, upper secondary, both sexes

— Gdp per capita (current us$)

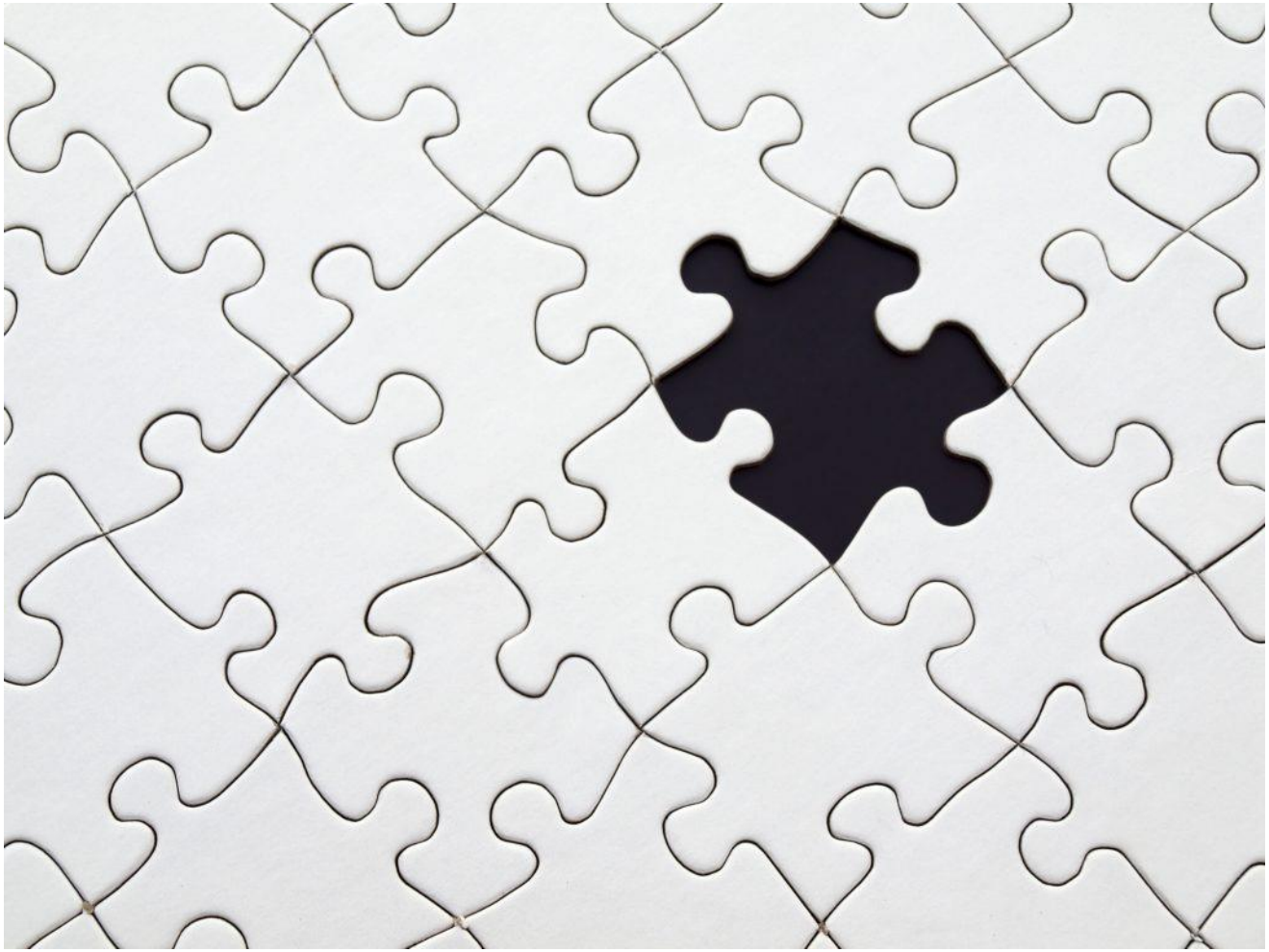
— Internet users (per 100 people)

— Population, ages 15-24, total

— population, total

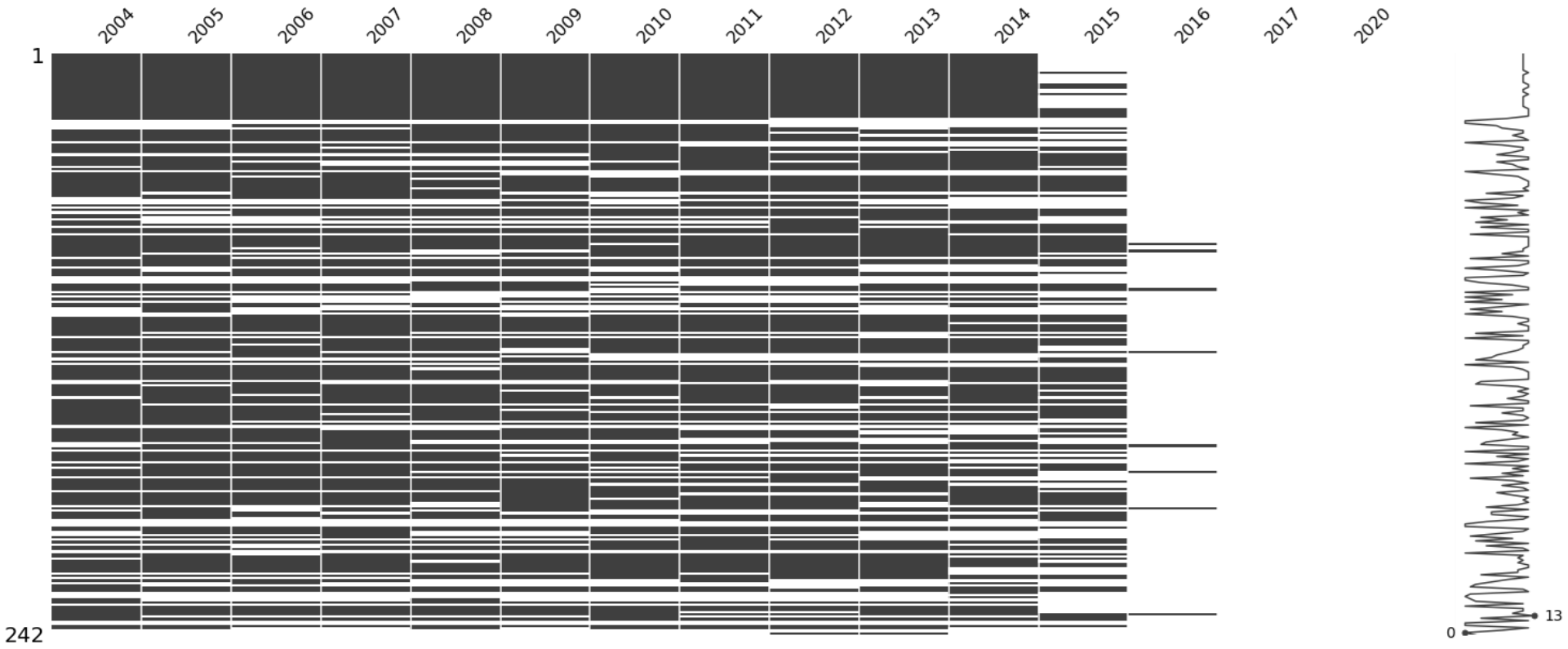
} Rapport des deux

IV) Qualité des données

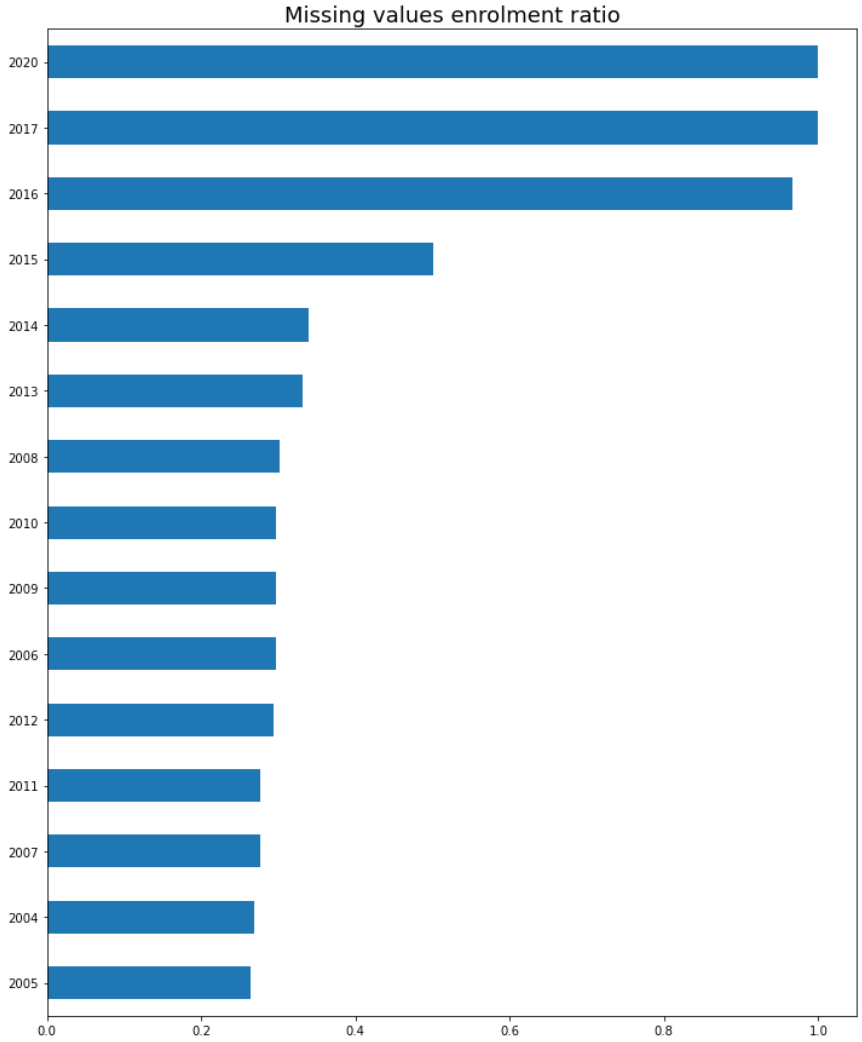


IV) 1- Scolarisation

Matrice missingno



IV) 1- Scolarisation



IV) 1- Scolarisation

Taille de l'échantillon :

Avec 100 % de remplissage : 54

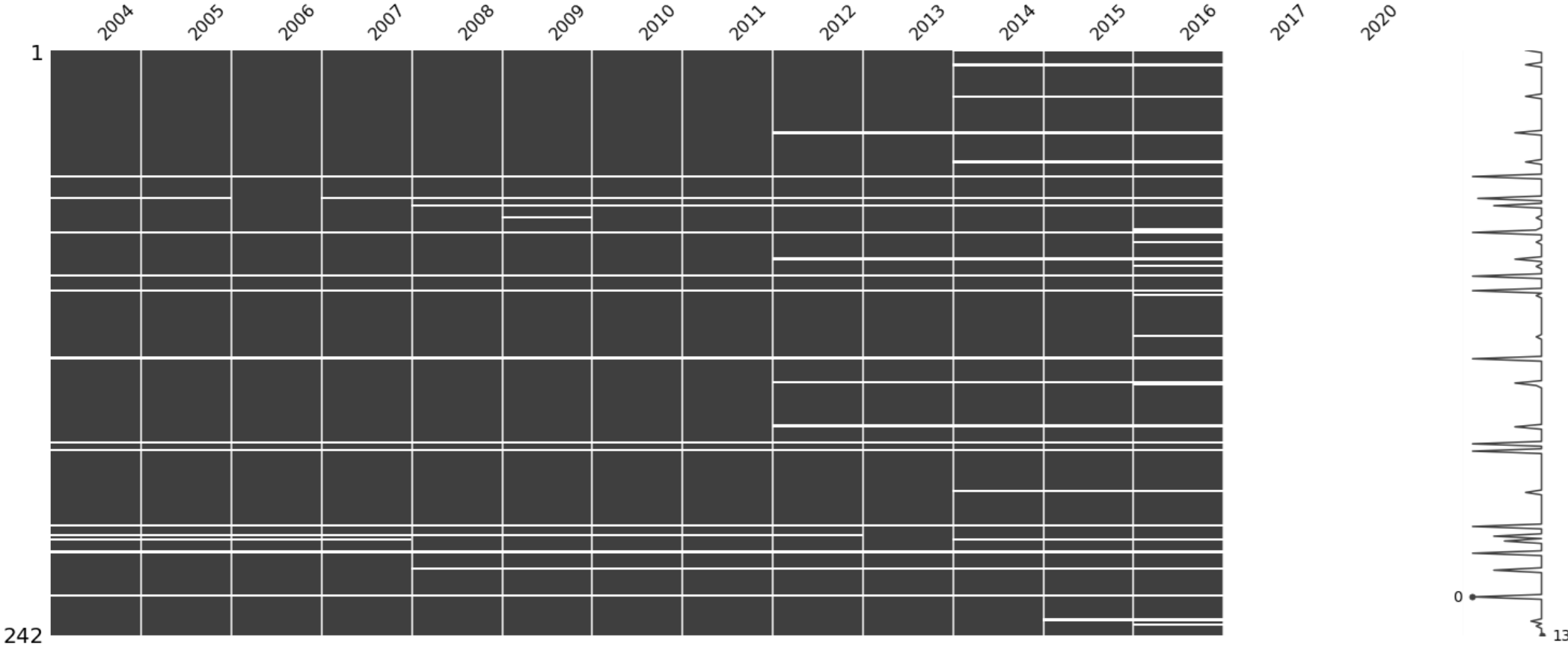
Avec 80 % de remplissage : 159

Bornes : [3.4, 220.3]

Pas de valeurs négatives ou non flottantes

IV) 2- PIB/hab

Matrice missingno



IV) 2- PIB/hab

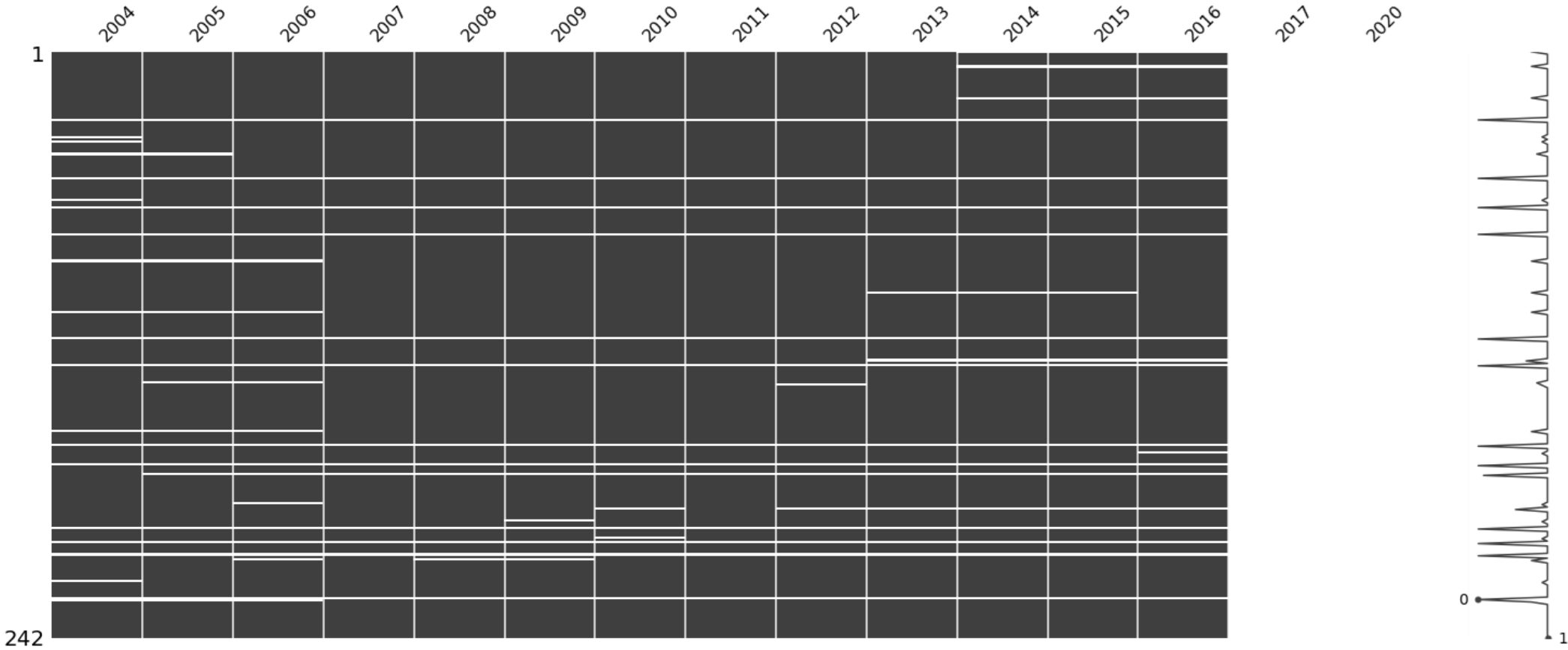
Taille de l'échantillon : 217

Bornes : [127.4, 179308.1]

Pas de valeurs négatives ou non flottantes

IV) 3- Internet

Matrice missingno



IV) 3- Internet

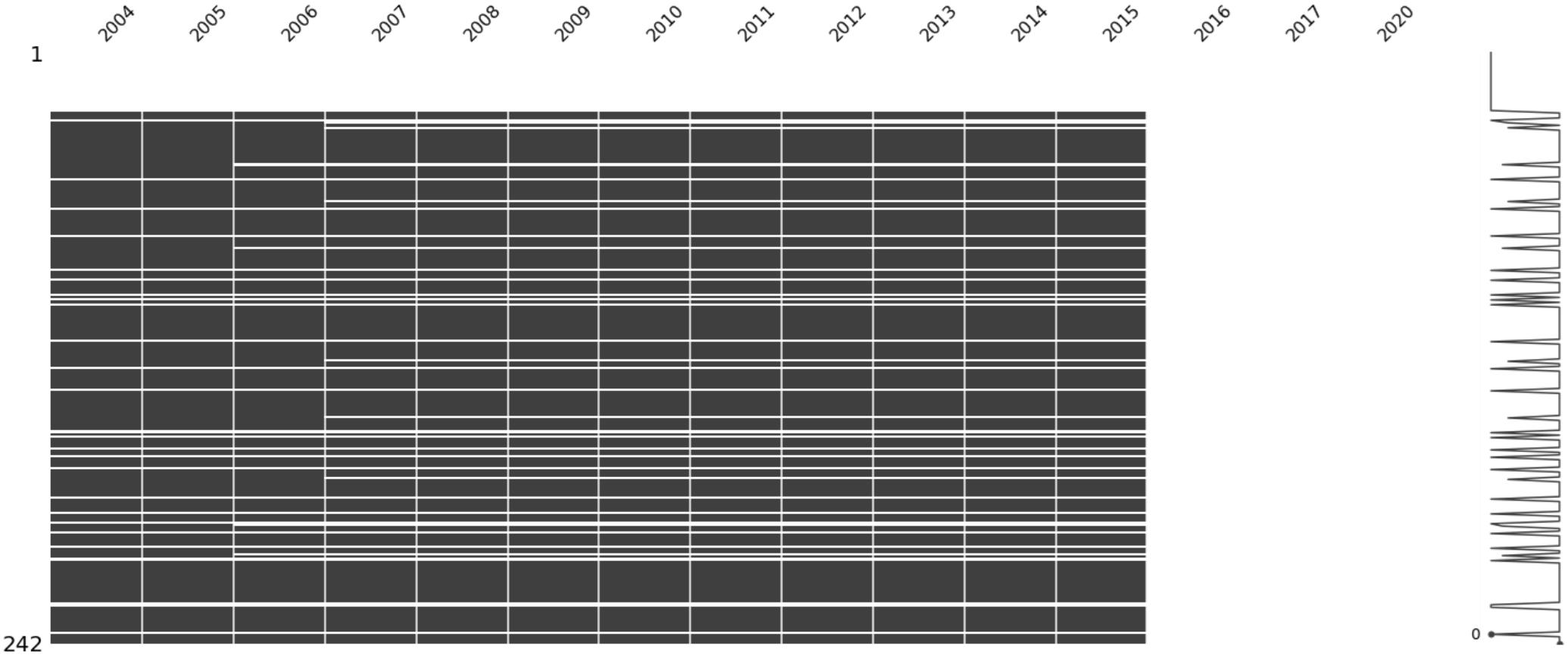
Taille de l'échantillon : 208

Bornes : [0.02, 98.16]

Pas de valeurs négatives ou non flottantes

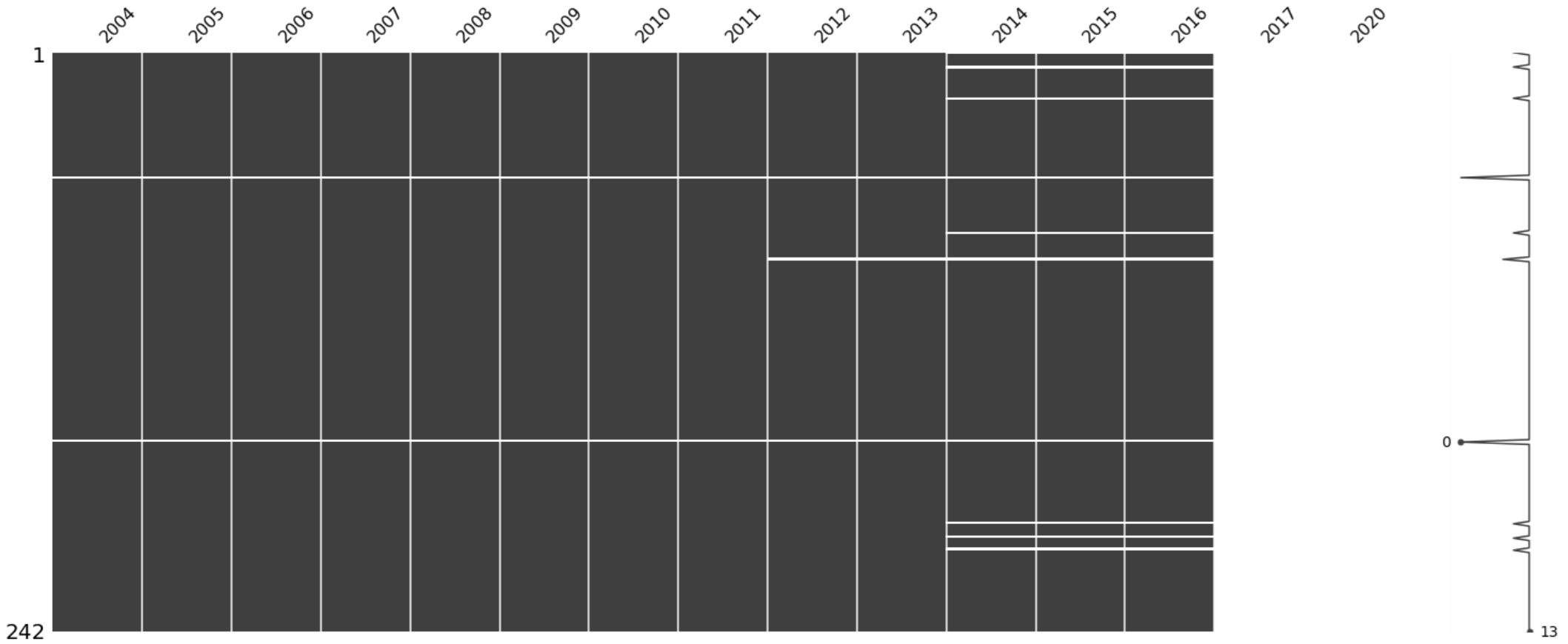
IV) 4- Démographie

Matrice missingno - Jeune



IV) 4- Démographie

Matrice missingno - Totale



IV) 4- Démographie

Taille de l'échantillon :

– Jeune : 181

– Totale : 232

Bornes du rapport : [0.05, 0.27]

Pas de valeurs négatives ou non flottantes

V) Corrélations

Avec le coefficient de Pearson :

Scolarisation – PIB : 0.54

Scolartisation – Démographie : -0.67

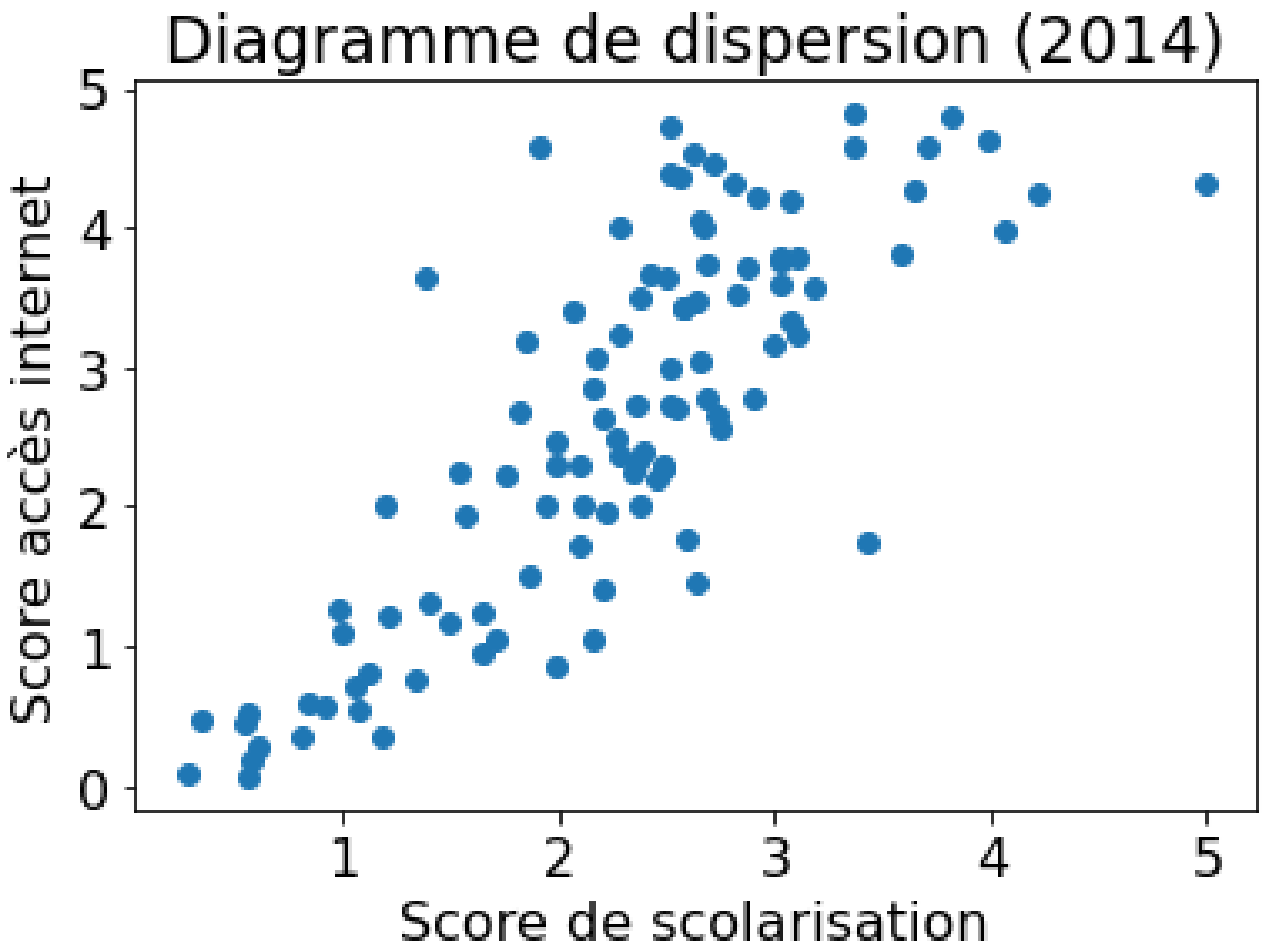
Scolarisation – Accès internet : 0.81

PIB – Démographie : -0.64

PIB - Accès internet : 0.75

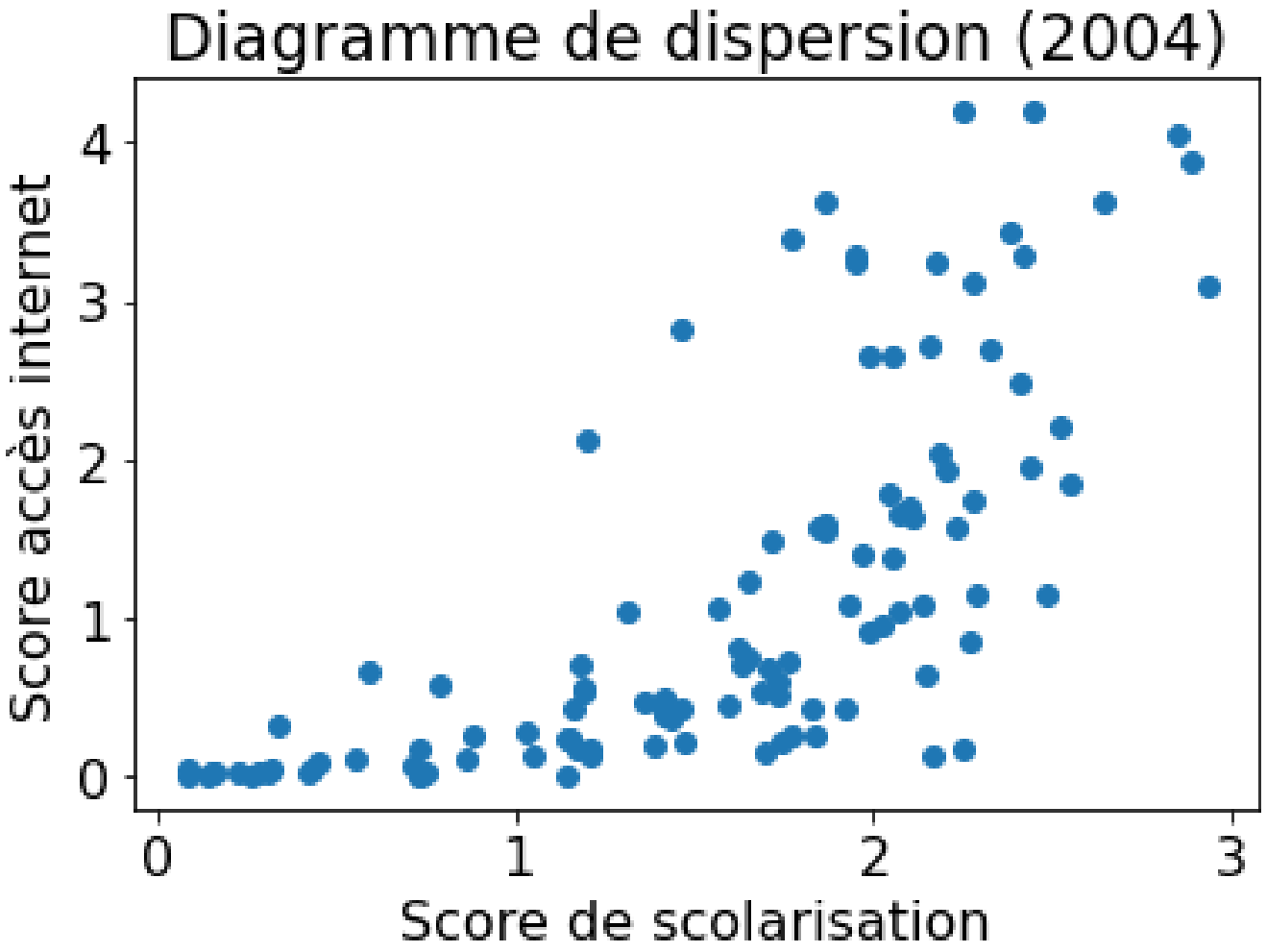
PIB – Démographie : -0.63

V) Corrélations



Coef Pearson : 0.81

V) Corrélations



Coef Pearson : 0.71

VI) Calcul du score

1- Scolarité : 5*ratio/ratio_max

2- Pib/hab : 5*pib/pib_max

3- Internet : 5*taux

4- Démographie : 5*taux/taux_max

Score total /20 : Somme_scores /5

Échantillon final : 119 pays

VII) Analyse des données par pays

Indicateurs étudiés :

– Moyenne → Sensible aux outliers

Sélection des valeurs entre Q1-1.5.EIQ et Q3+1.5.EIQ

Méthodes : quantile, df.index.map

– Médiane

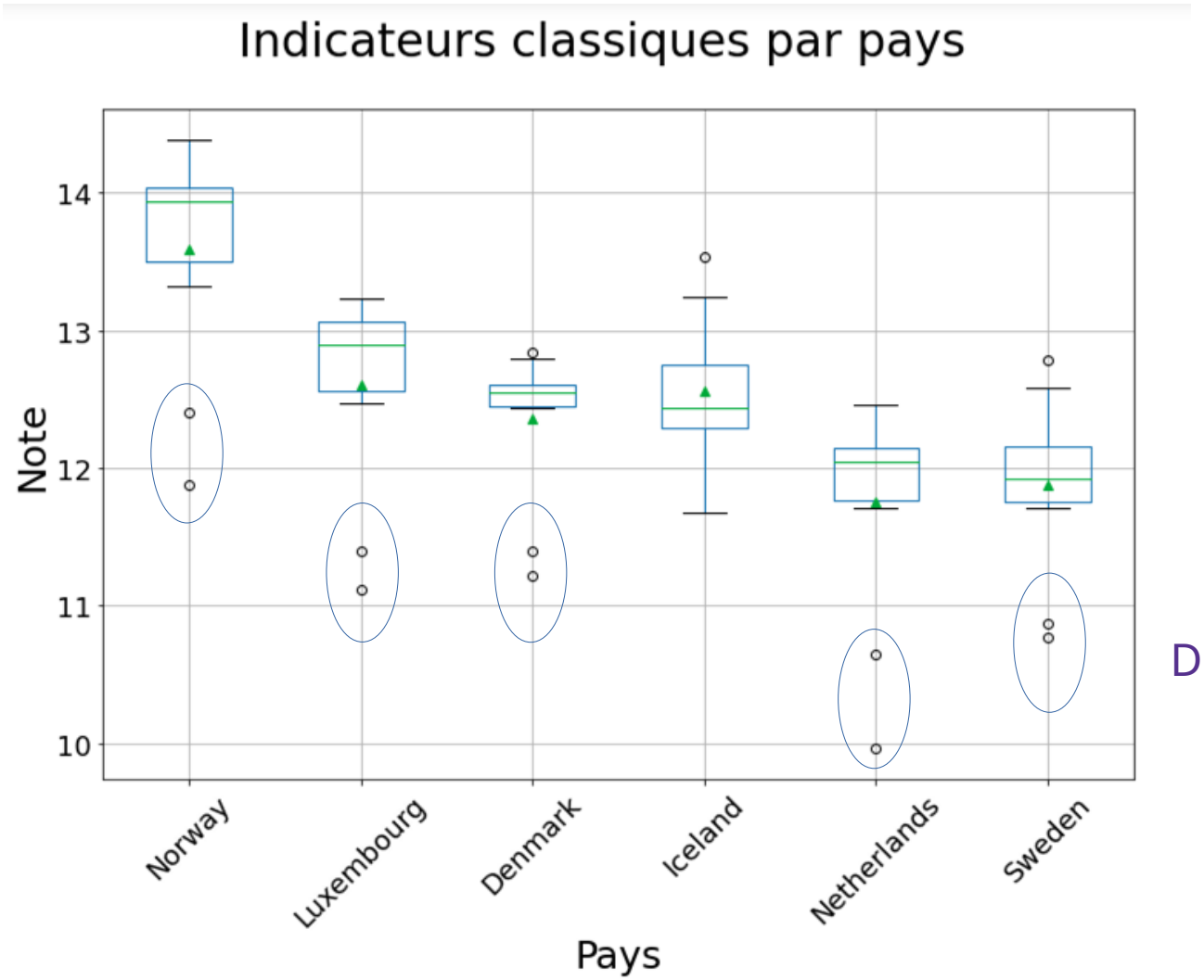
– Écart-type

VII) 1- Pays au plus fort potentiel

On classe les pays avec leur score moyen :

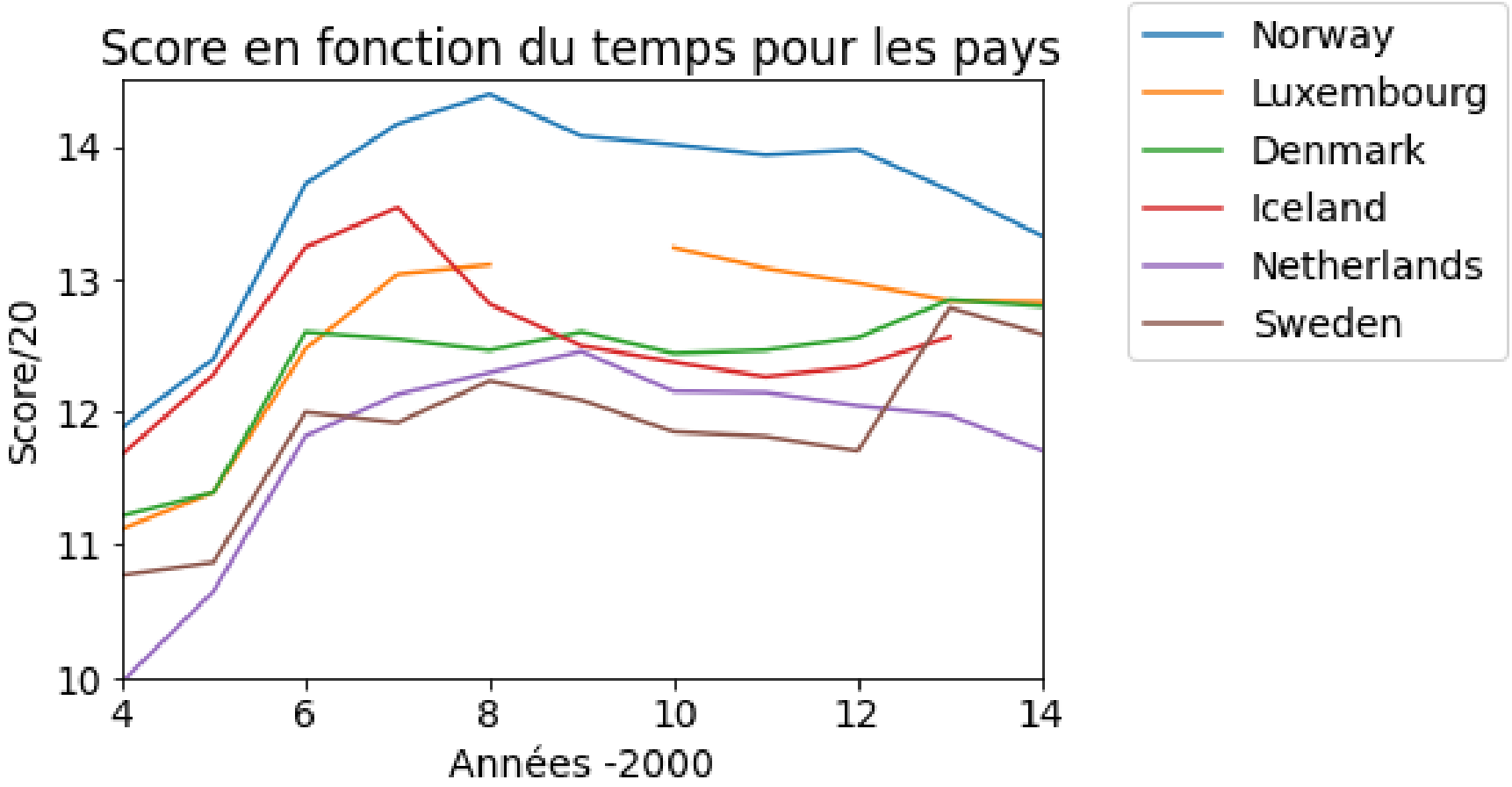
Country Name	Mean	Median	Standar deviation
Norway	13.917501	13.932526	0.780389
Luxembourg	12.945753	12.901924	0.744648
Denmark	12.559033	12.547185	0.536906
Iceland	12.450190	12.435768	0.528735
Netherlands	12.080895	12.045138	0.762123
Sweden	12.023910	11.918853	0.615509

VII) 1- Boîtes de Tukey

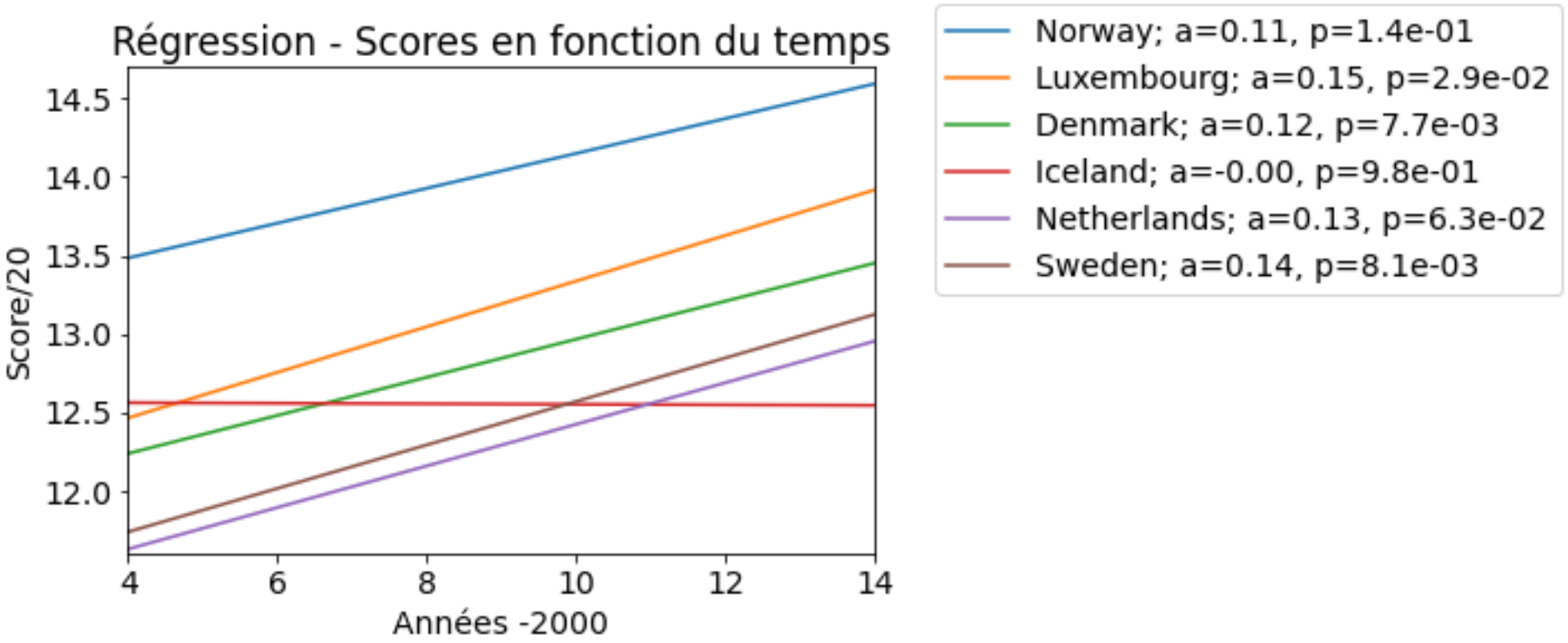


Données anciennes

VII) 1- Courbes des scores



VII) 2-Évolution du potentiel



VII) 2- Taux de croissance des scores

Calcul du taux de croissance moyen entre 2004 et 2014

	Mean growth	Country Name	
628090	0.011558	Norway	
514475	0.014409	Luxembourg	←
287245	0.013248	Denmark	
411855	0.000000	Iceland	
602435	0.016281	Netherlands	←
778355	0.015686	Sweden	←

VII) 2- Analyse des données prospectives

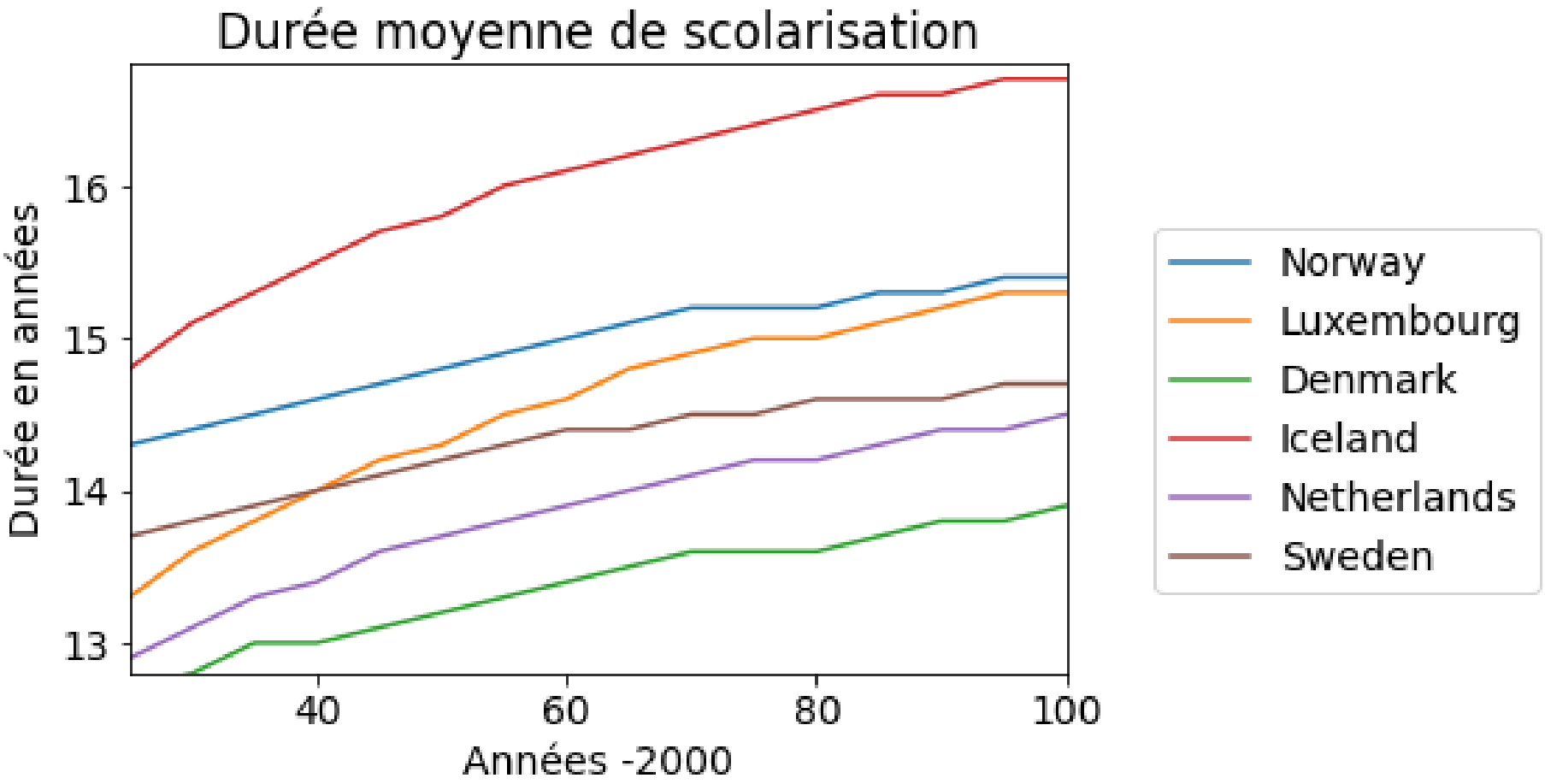
Indicateurs de l'analyse rétrospective : Aucune donnée disponible

Nouveaux indicateurs choisis :

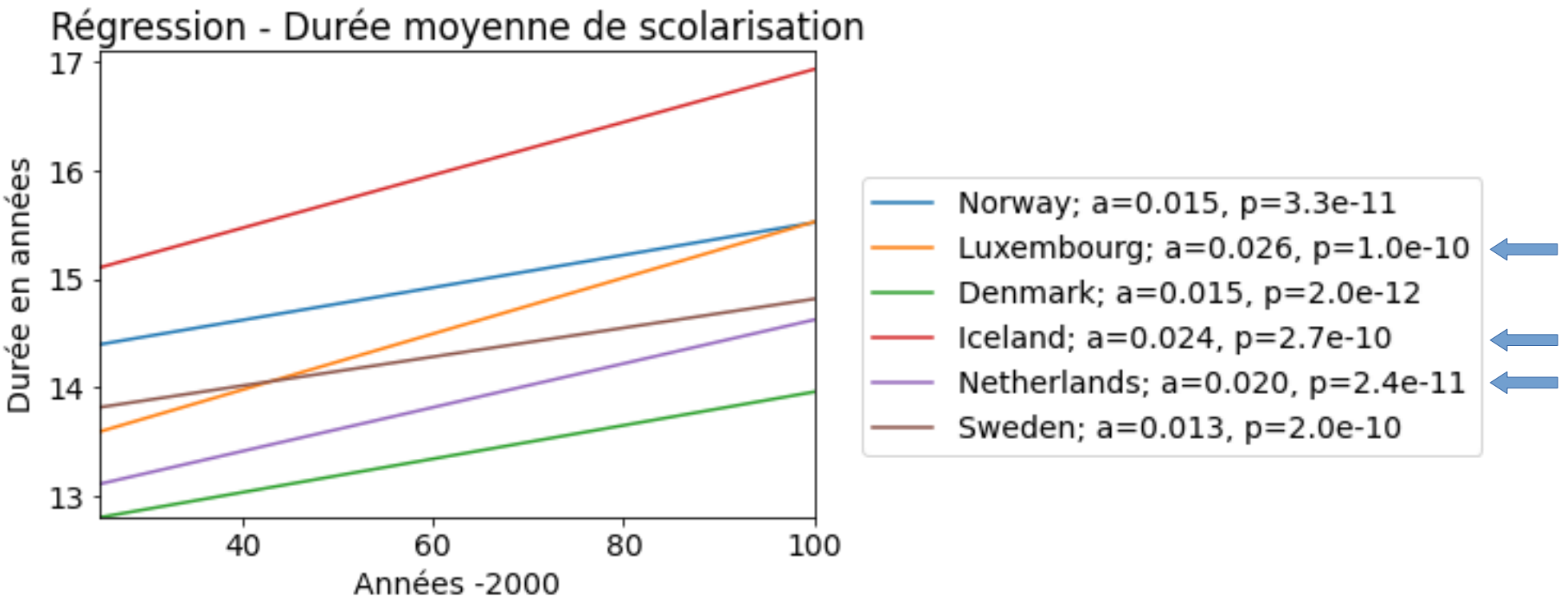
- projection: mean years of schooling. Age 20-39

- Population by country and region, historic and projections (depuis *ourworldindata*)

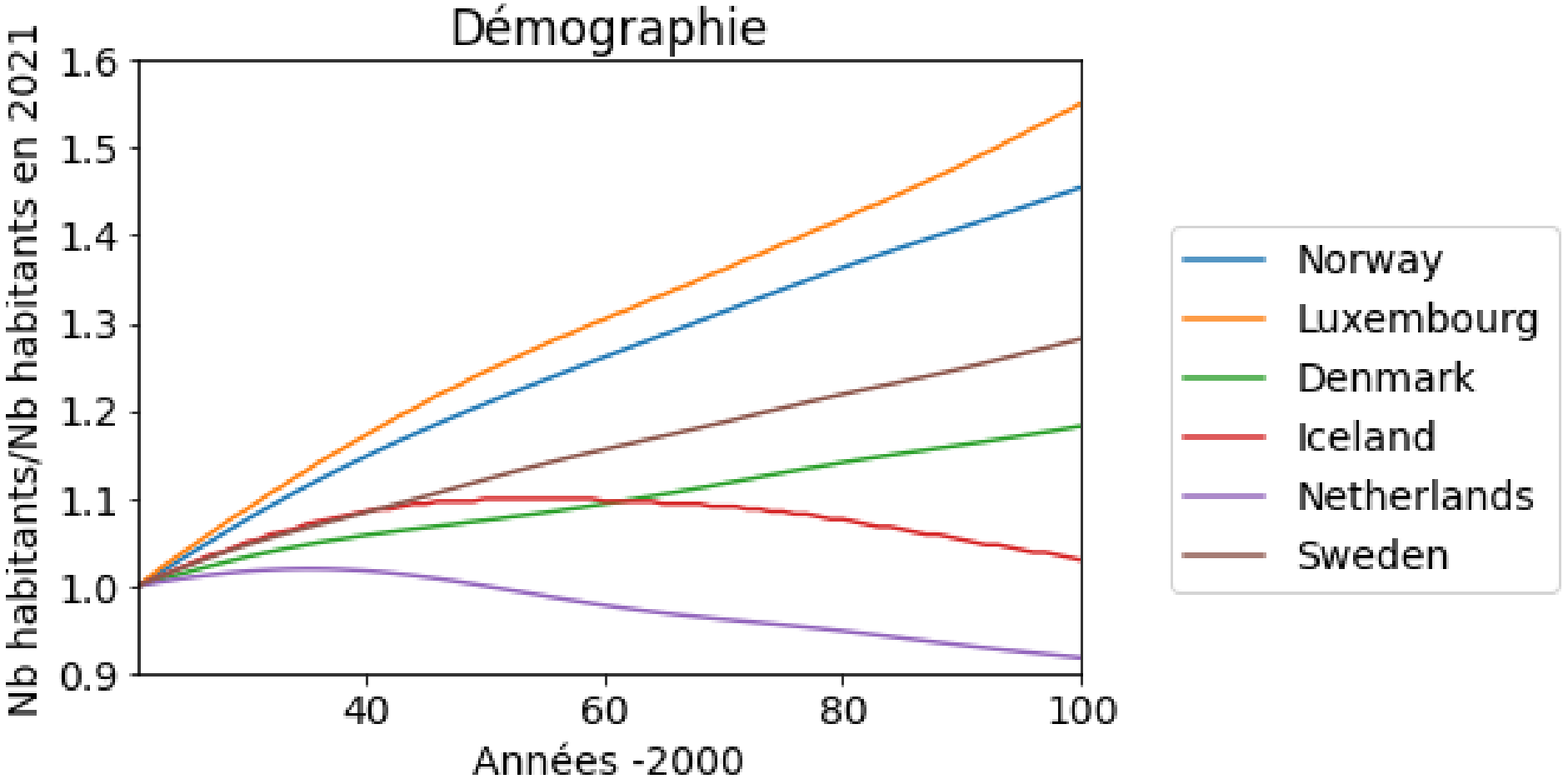
VII) 2- Durée moyenne de scolarisation



VII) 2- Durée moyenne de scolarisation



VII) 2- Démographie



VIII) Pays à investir en priorité

Taux de croissance

1– Netherlands +3pts
2- Sweden +2pts
3- Luxembourg +1pts

Score moyen

1– Norway
2- Luxembourg
3- Denmark

Durée de scolarisation

1– Luxembourg
2- Iceland
3- Netherlands

Démographie

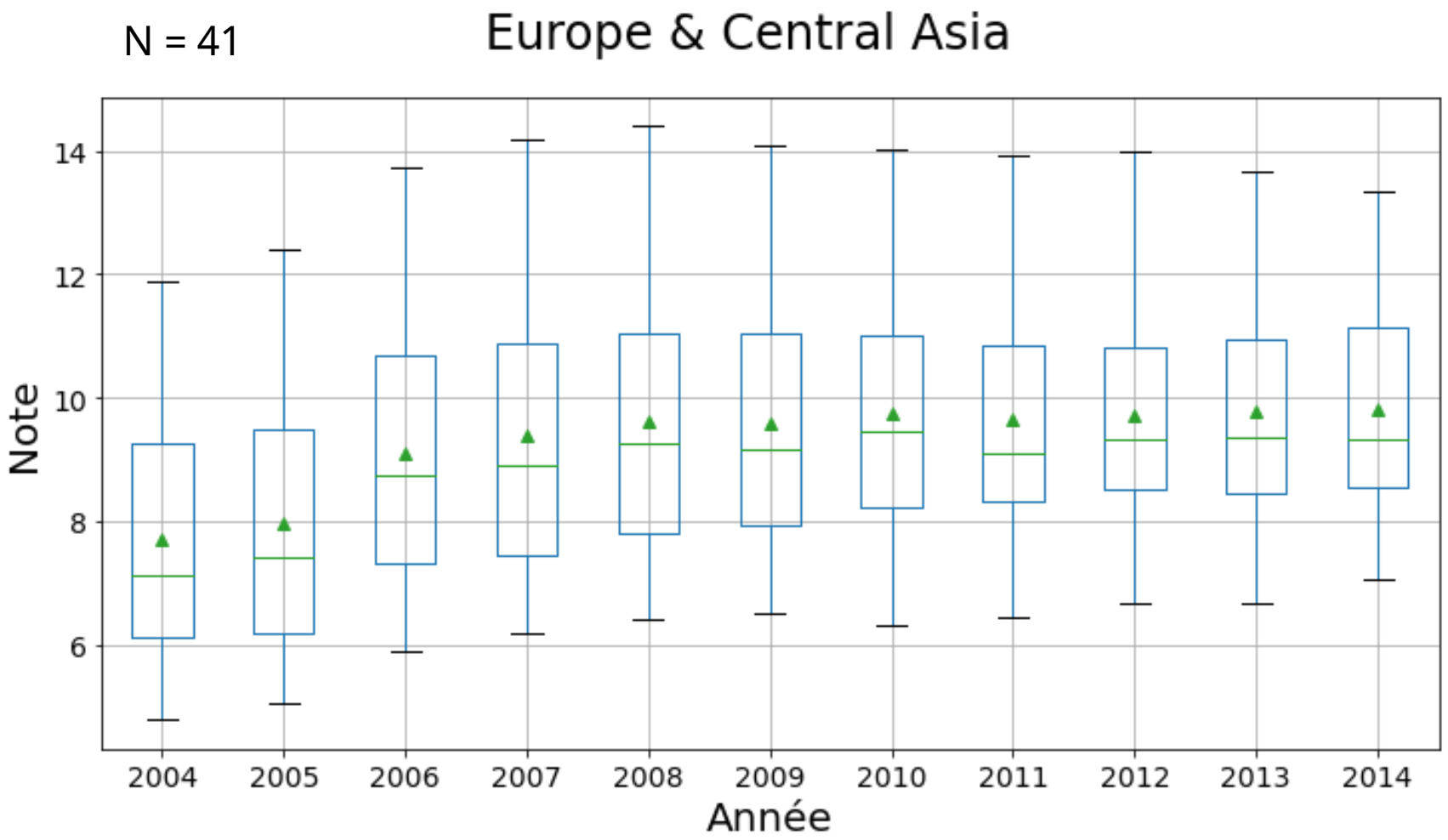
1– Luxembourg
2- Norway
3- Sweden

Score de priorité

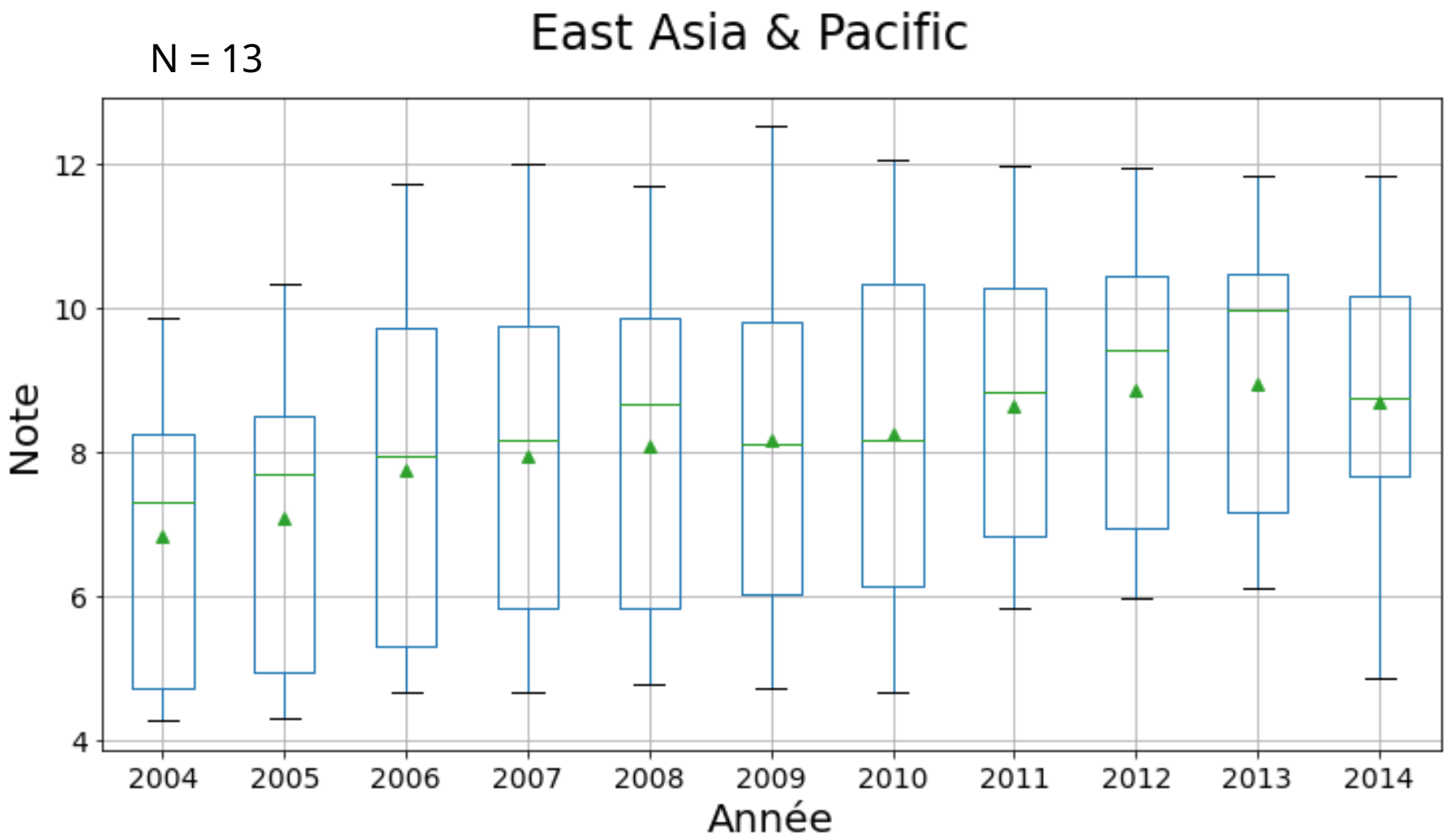
Luxembourg : 14
Norway : 5
Netherlands : 4
Sweden : 3
Iceland : 2
Denmark : 1

Pays à investir en priorité

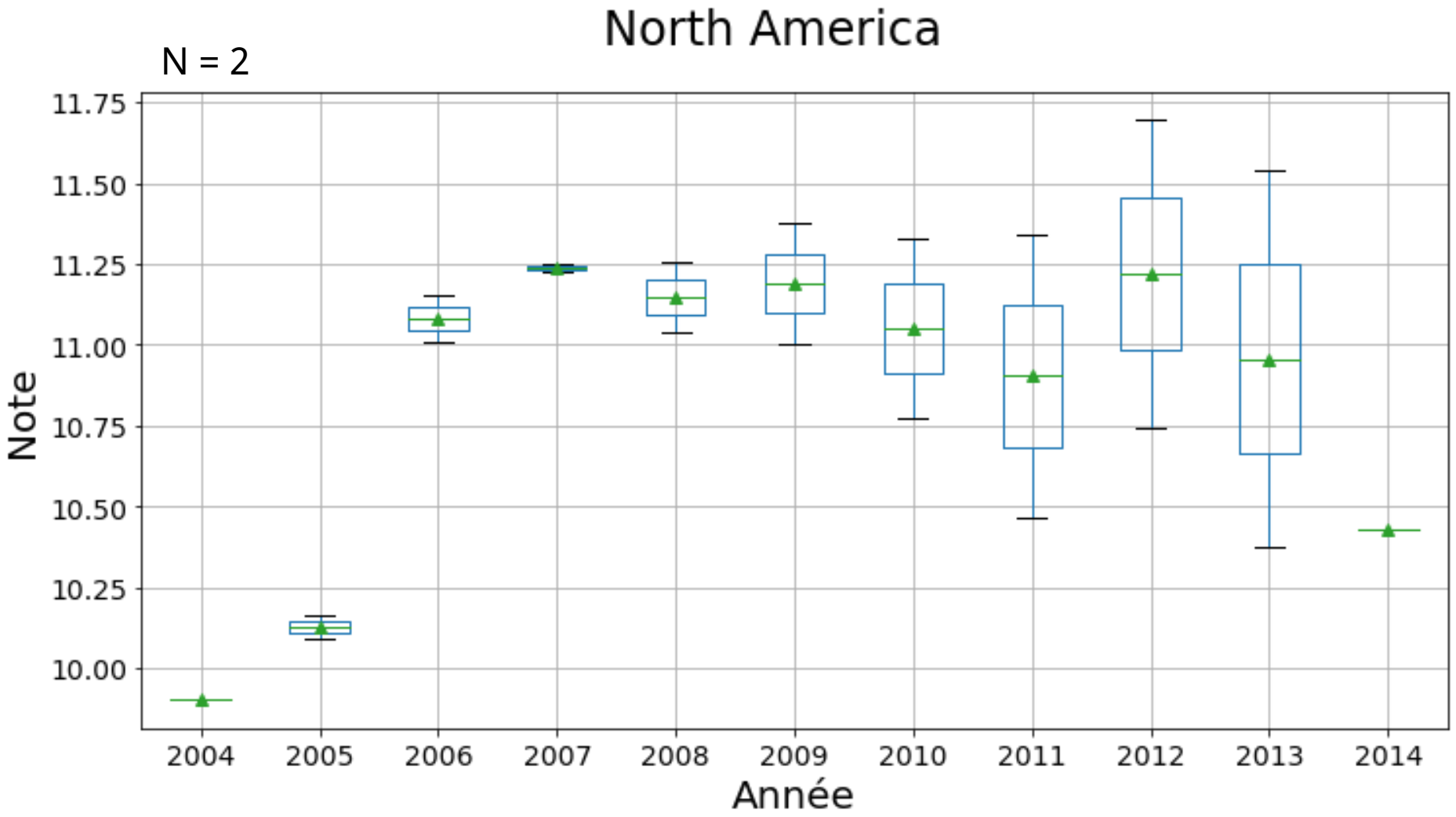
IX) Analyse des données par région



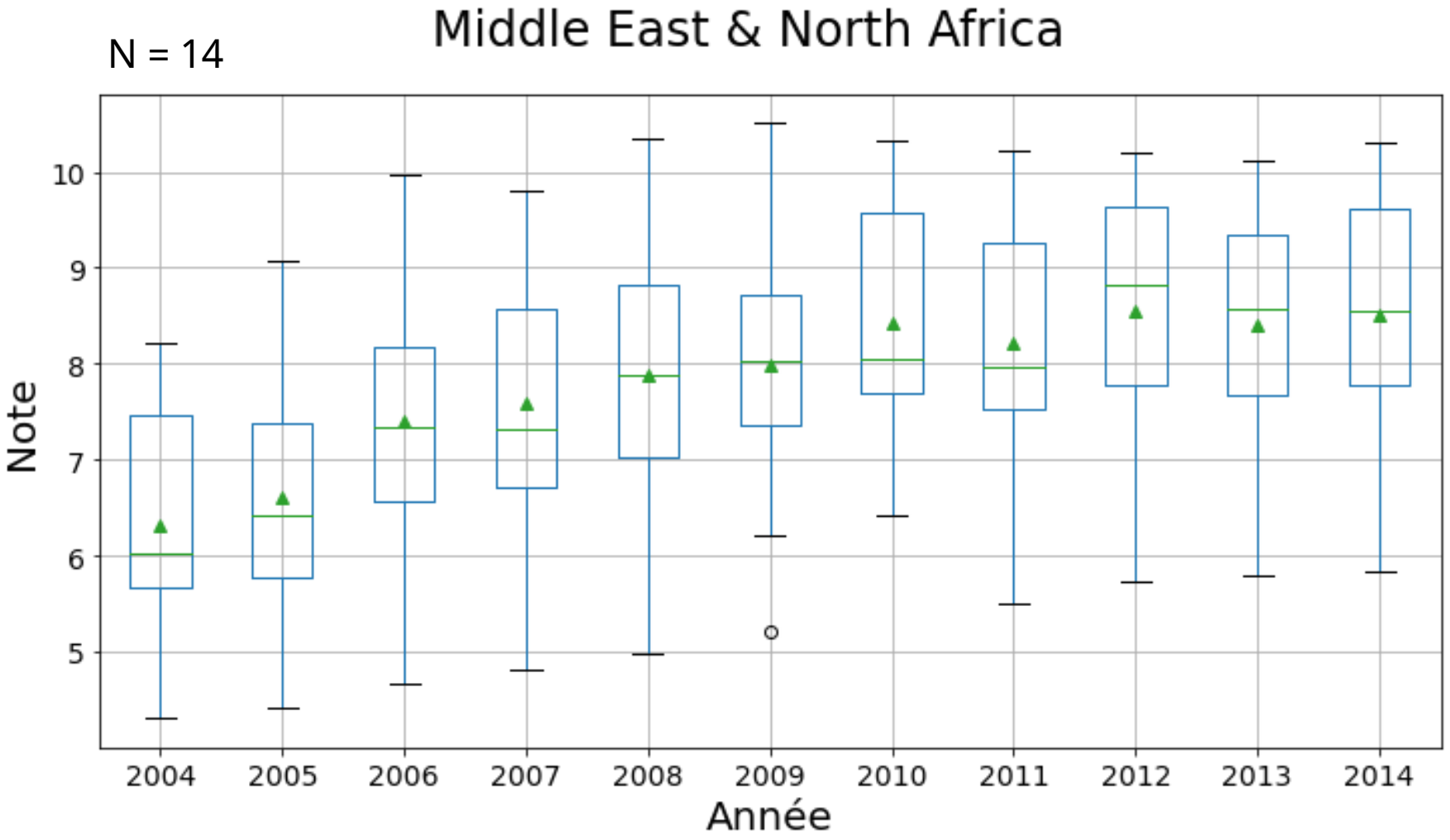
IX) Analyse des données par région



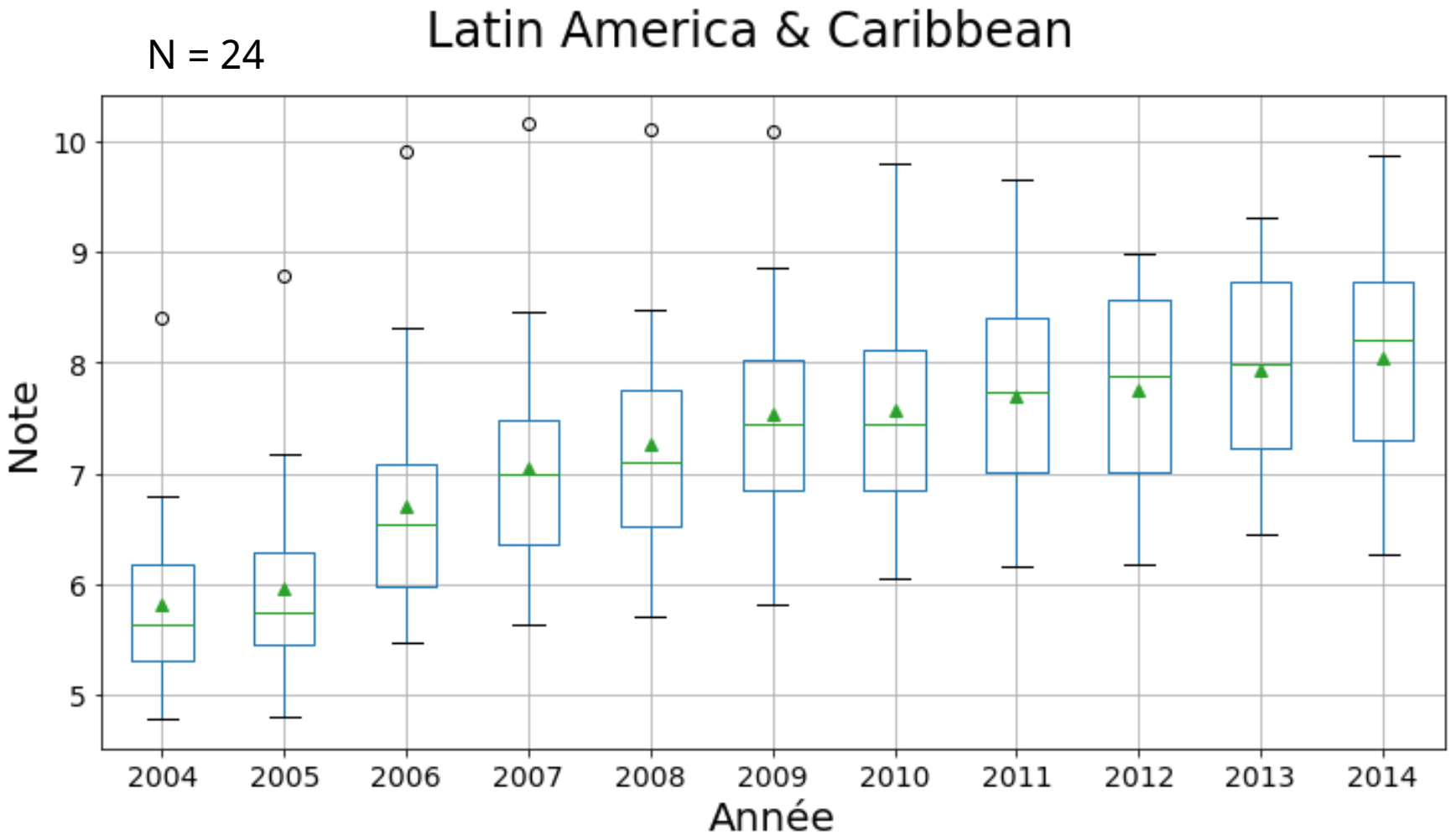
IX) Analyse des données par région



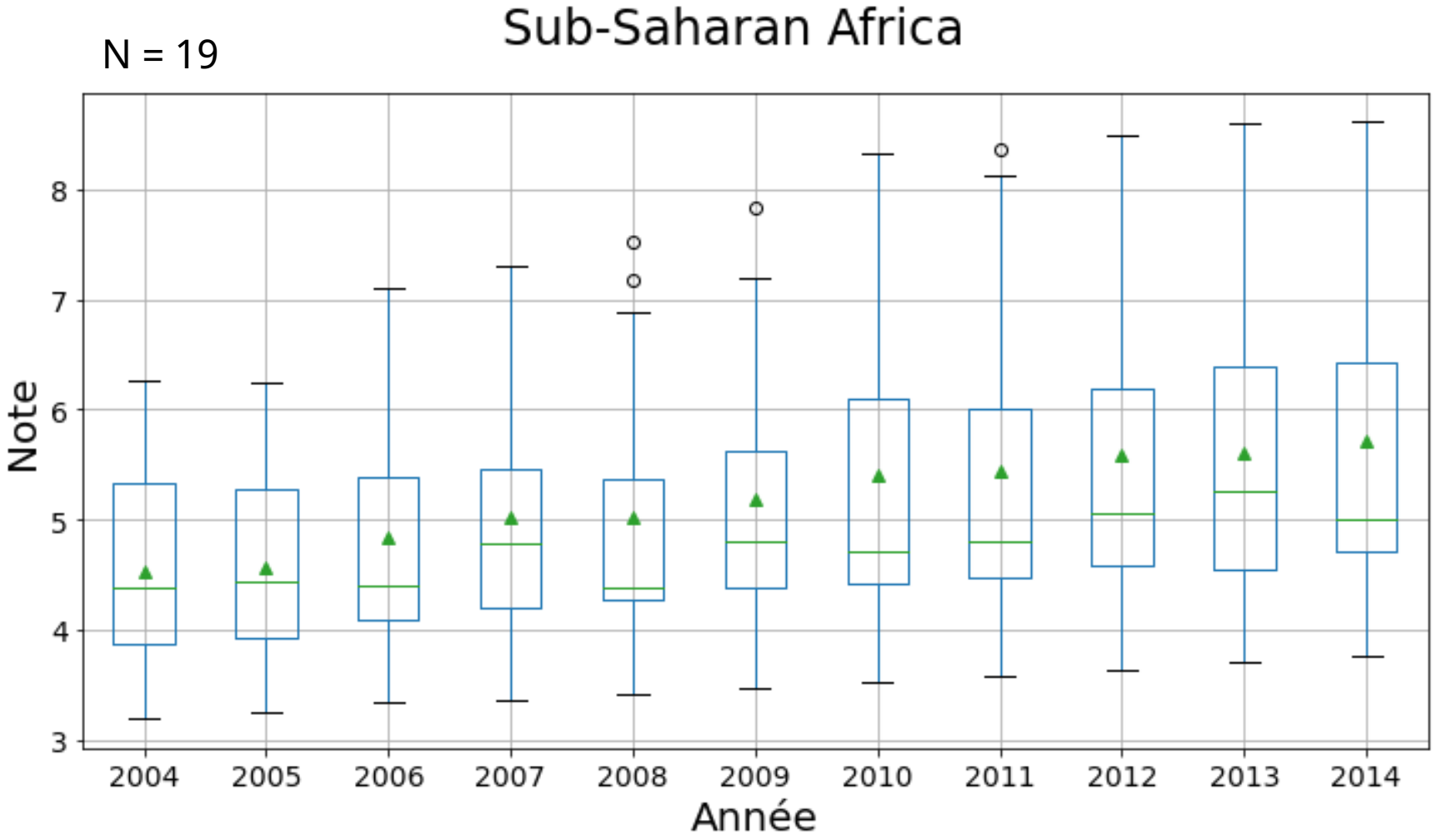
IX) Analyse des données par région



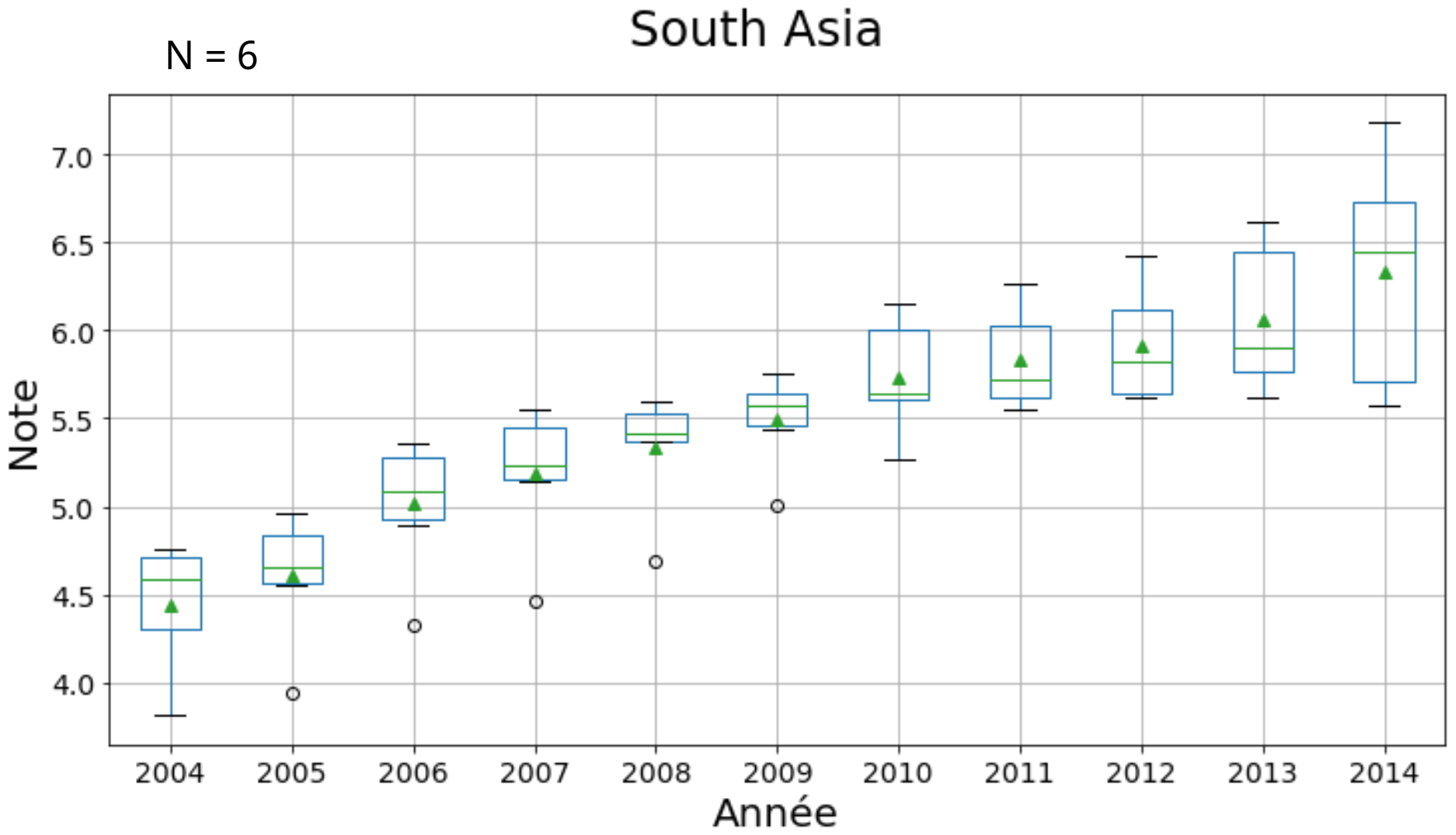
IX) Analyse des données par région



IX) Analyse des données par région



IX) Analyse des données par région



IX) Analyse des données par région

Écart-type des notes par région

Region		N
North America	0.343852	2
South Asia	0.385965	6
Latin America & Caribbean	0.941495	24
Sub-Saharan Africa	1.226545	19
Middle East & North Africa	1.335373	14
Europe & Central Asia	1.959742	41
East Asia & Pacific	2.226464	13

dtype: float64

Conclusion

Les pays à investir en priorité sont, dans l'ordre, le Luxembourg, la Norvège et le Danemark.

D'autres indicateurs auraient pu permettre d'affiner les scores.

Avoir les données prospectives des indicateurs initiaux aurait permis une évaluation plus précise de l'évolution du score.