Alexandre ROGUES

Parcours Data Scientist

Projet 2

Analyse de données de systèmes éducatifs

Janvier 2024

Analyse exploratoire

Développement COMMERCIAL *à l'INTERNATIONAL*

Alexandre ROGUES - Département Data - Janvier 2024

Nos objectifs

- offrir nos services aux delà des frontières
- s'assurer du succès de nos formules
- contribuer à l'éducation

Explorer les potentiels de chaque pays

EdStats All Indicator Query

Un dataset de la Banque Mondiale

- 4000 indicateurs
- + 200 pays
- + 60 années

Jupyter Lab Notebook

Outil python

- Environnement interactif
- Bibliothèques Pandas, Matplotlib...
- Modules de visualisation

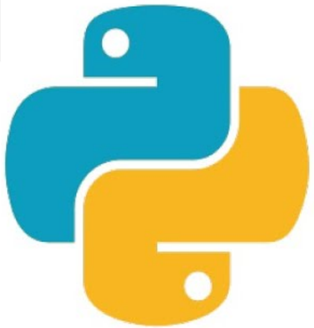
Jupyter Lab Notebook

Préparation du système

- Installation des libraires Python nécessaires
- Vérification des versions

```
pip install pandas
```

```
!pip freeze
```



Création d'un environnement virtuel

- Isoler les dépendances et packages Python spécifiques au projet
- Eviter les conflits entre différentes versions de bibliothèques
- Facilitation de la gestion des environnements de développement
- Reproductibilité garantie des analyses
- Gestion efficace des dépendances pour des projets Python

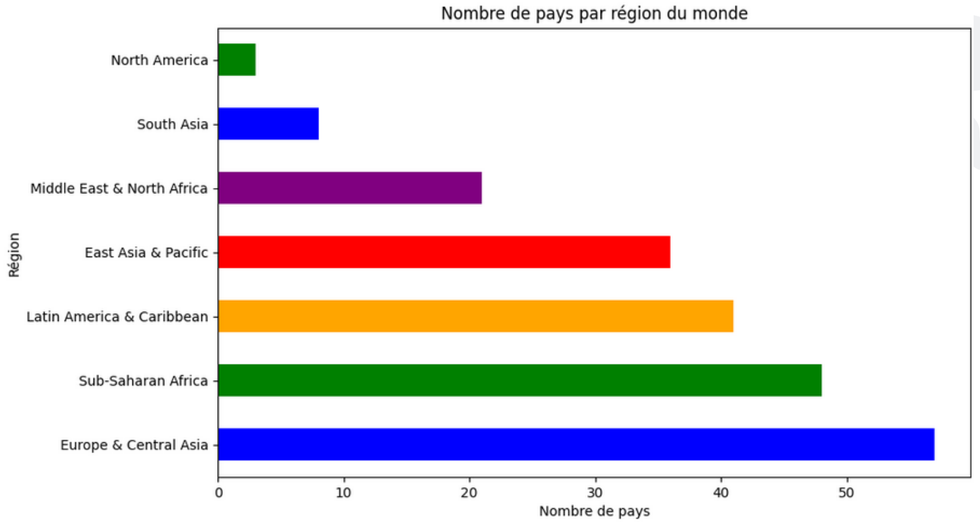
```
conda create --stabadenvP2
conda activate stabadenvP2
```

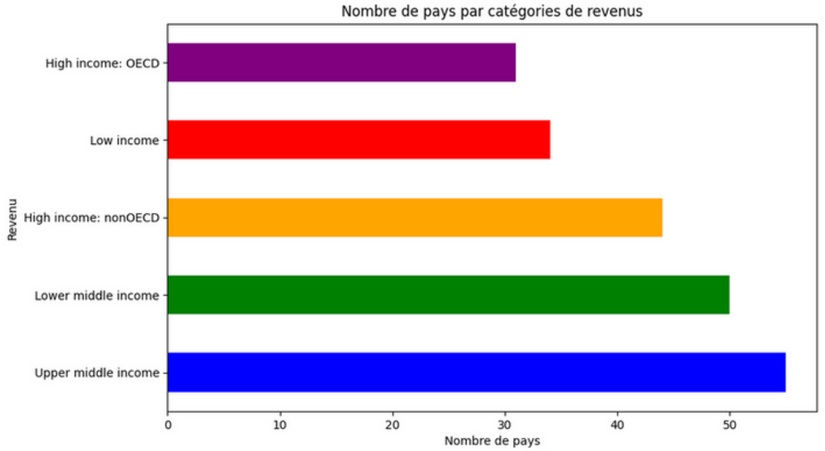
EdStats All Indicator Query

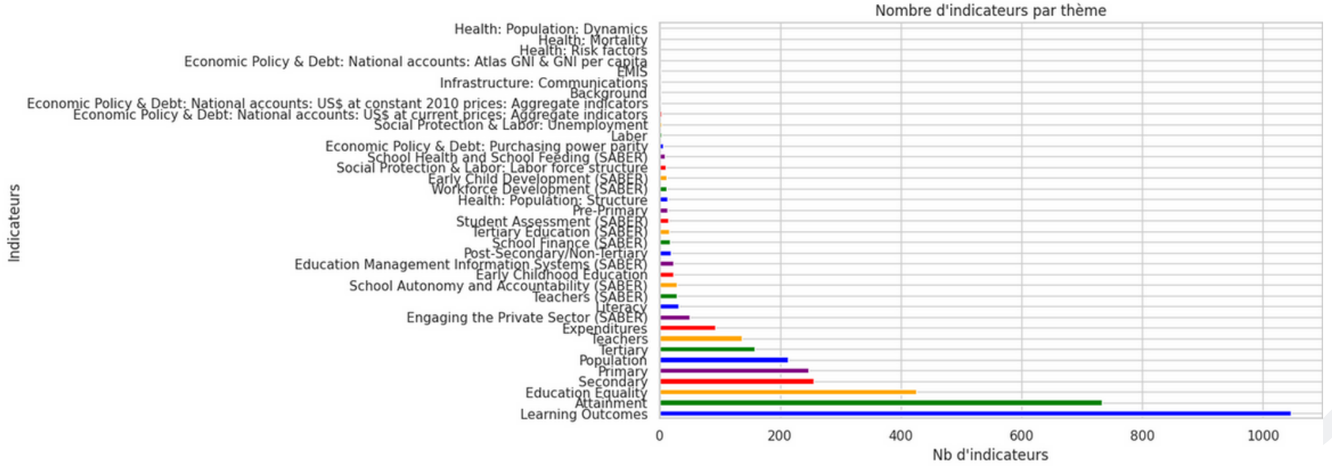
EdStatsCountry-Series.csv	47.8 KB
EdStatsCountry.csv	136.2 KB
EdStatsData.csv	311.3 MB
EdStatsFootNote.csv	37.9 MB
EdStatsSeries.csv	3.5 MB

5 fichiers csv

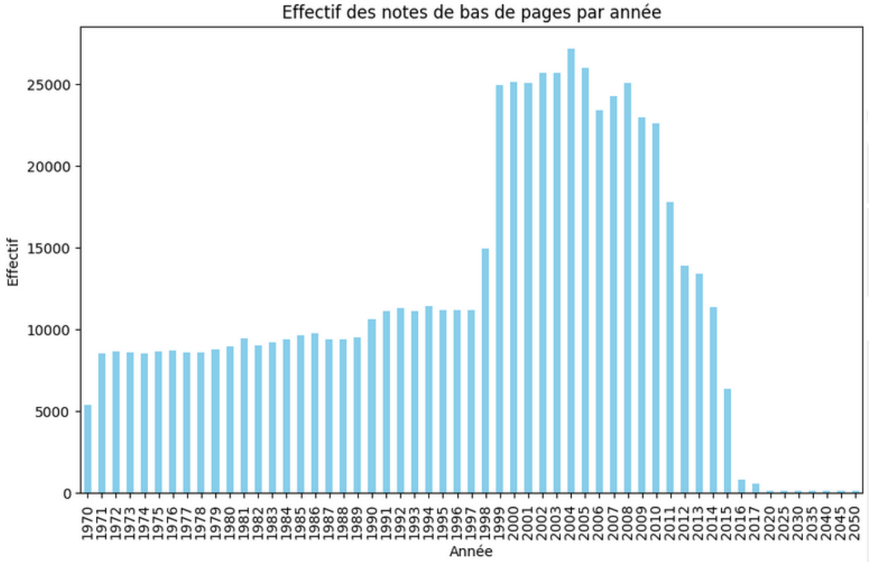
+300Mo data







EdStats All Indicator Query



Un taux de remplissage annuel inégal

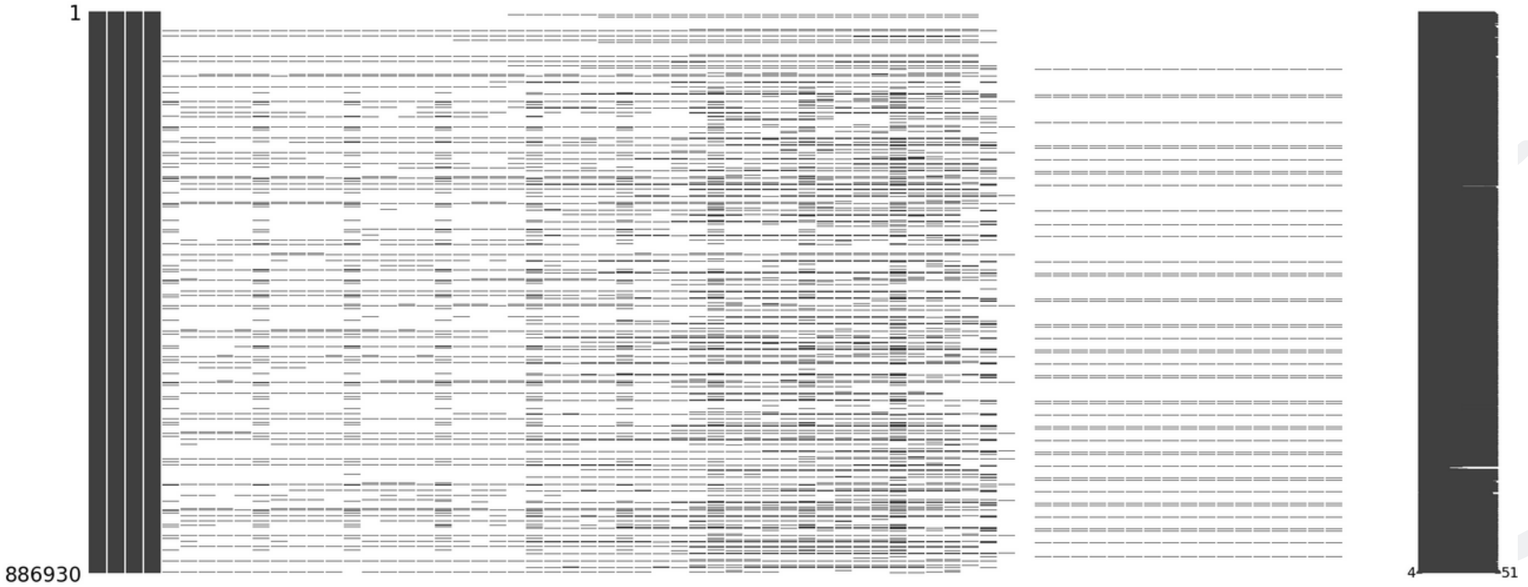
EdStatsData.csv

Un fichier lourd et néanmoins incomplet

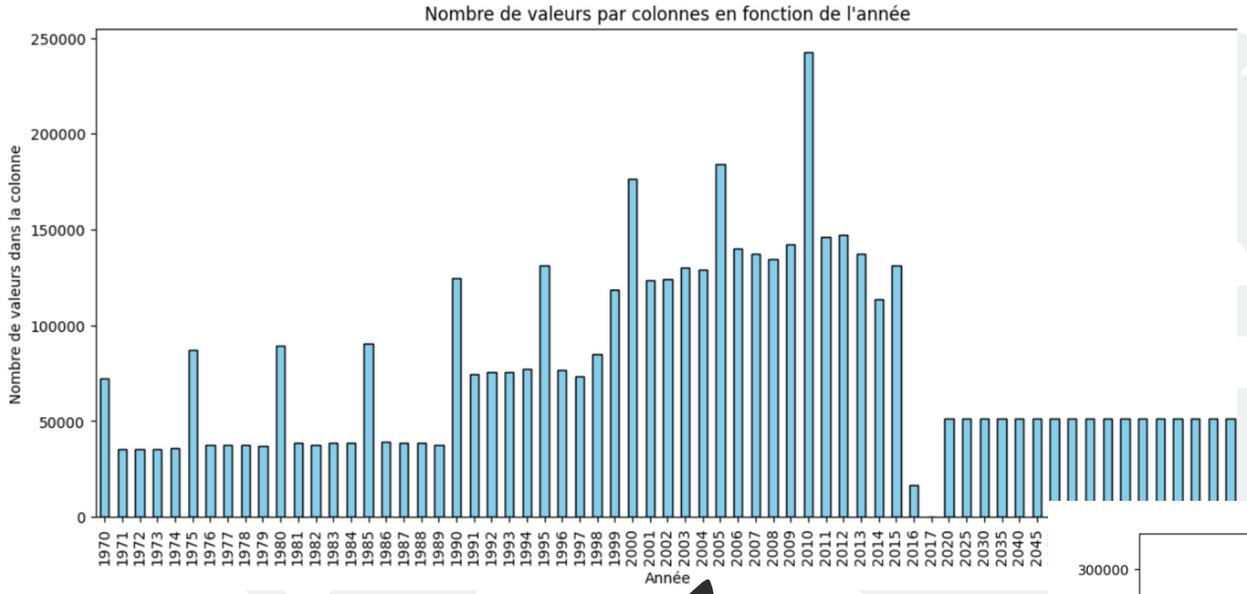
```
#chargement du fichier dans un dataframe pandas
df_data = pd.read_csv('EdStatsData.csv')

#structure du df
df_data.shape
```

```
(886930, 70)
```

```
ms.matrix(df_data)
```



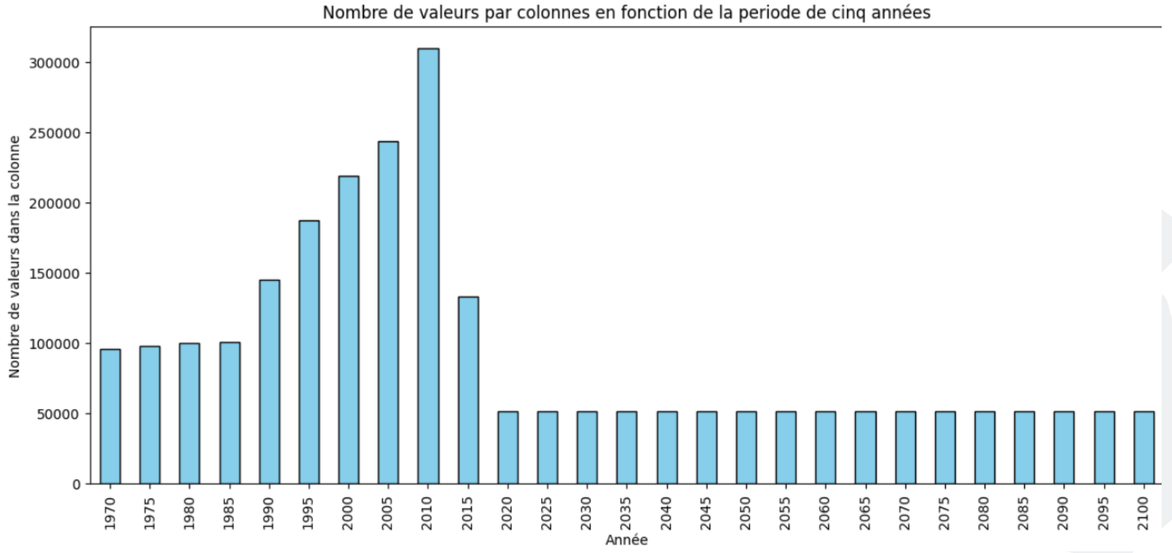
EdStatsData.csv



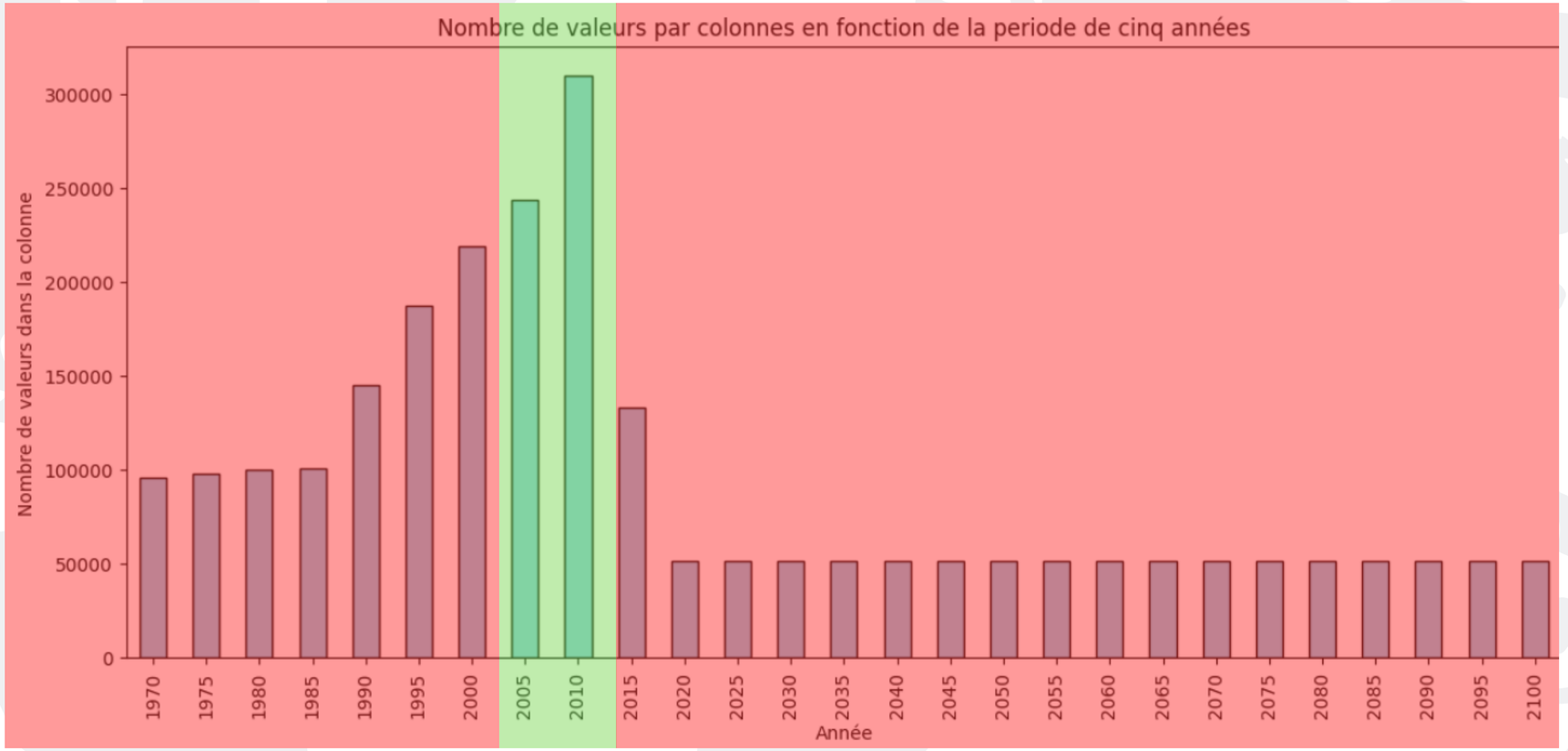
Stratégie 1

Exploration par nombre de données annuelles

Regroupement des données sur plage quinquennale



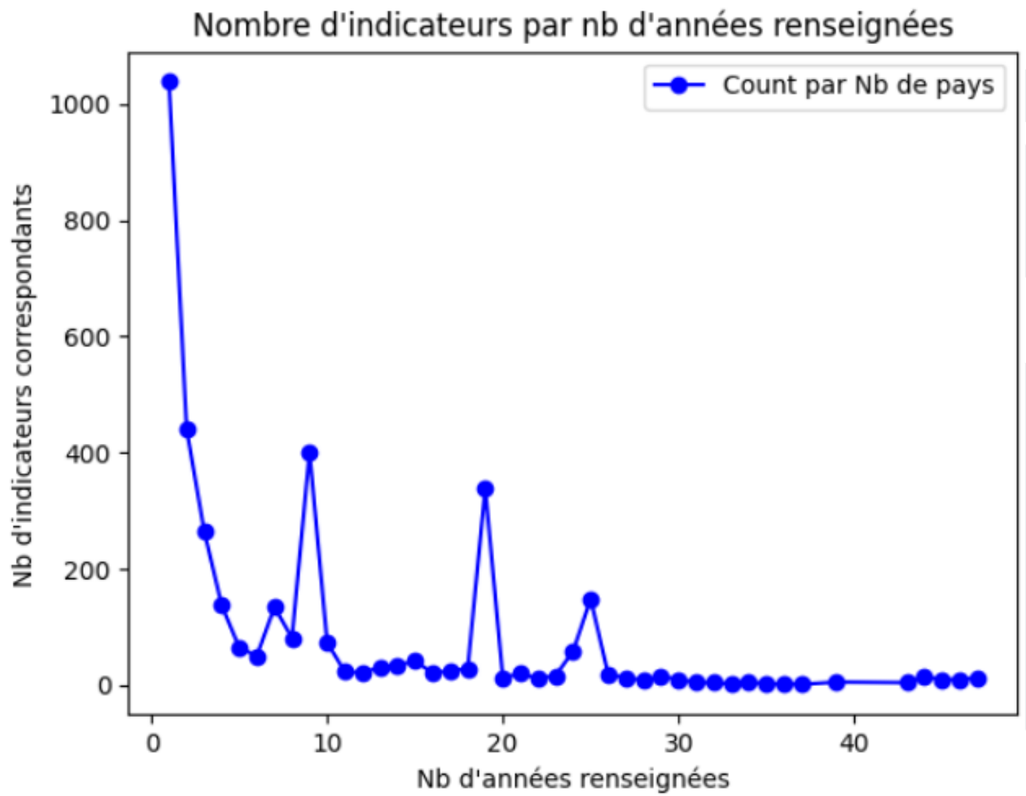
EdStatsData.csv

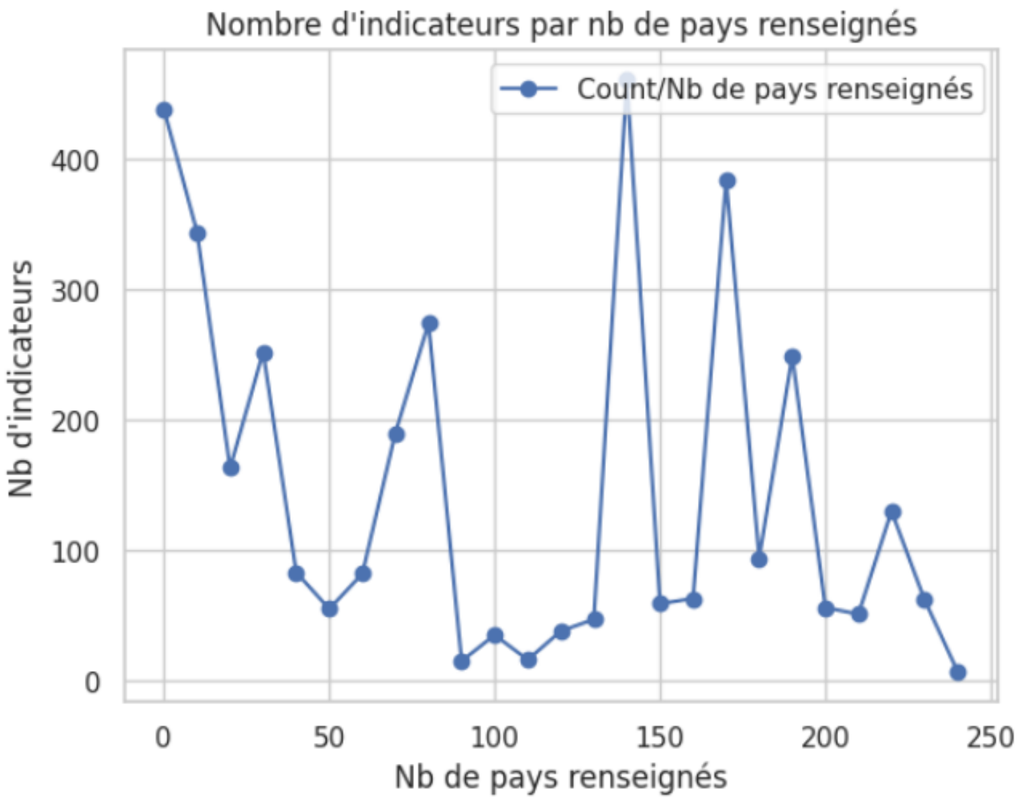


Deux périodes les plus renseignées

- 2005 - 2010
- 2010 - 2015

EdStatsData.csv

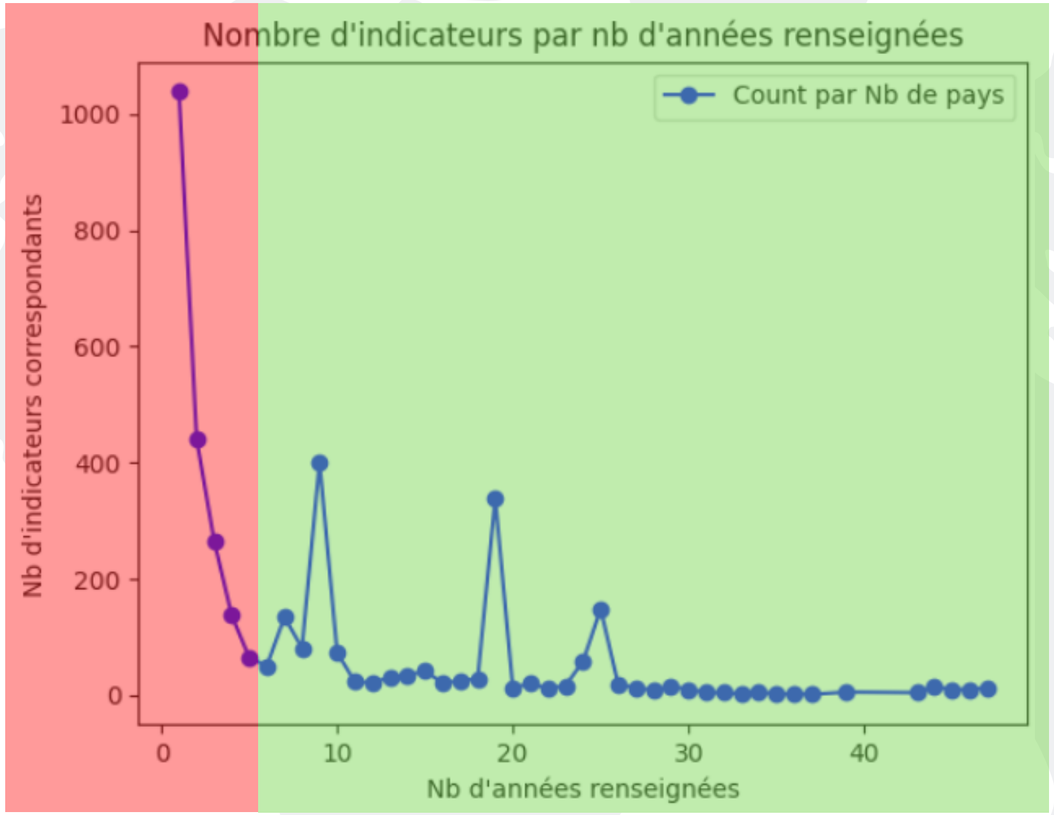


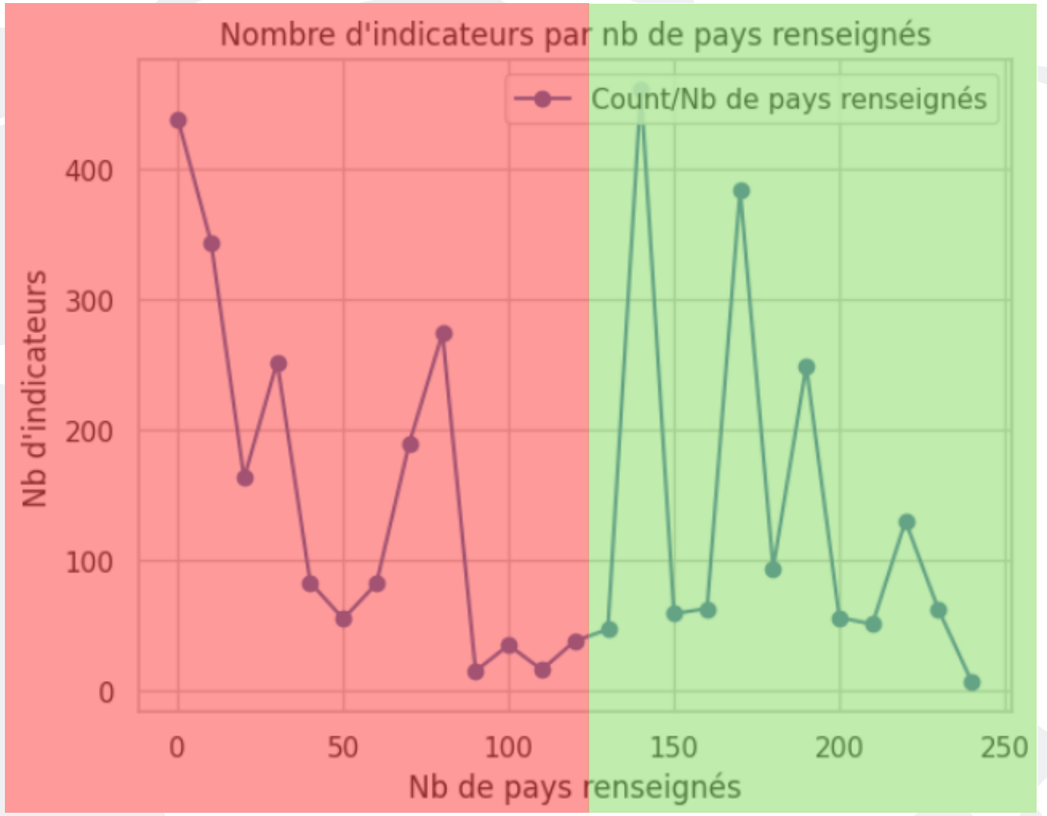


Stratégie 2

Exploration par données / indicateurs

EdStatsData.csv



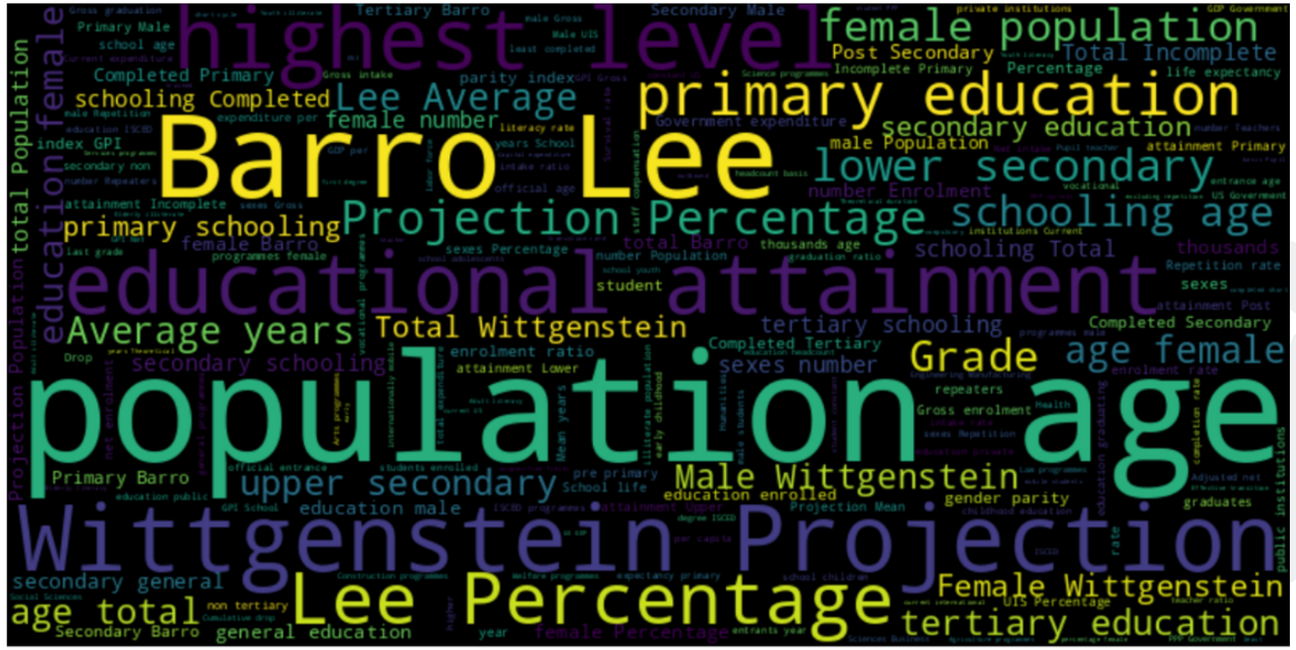


Sélection d'indicateurs renseignant à propos de

- + 4 années
- + de 125 pays

EdStatsData.csv

nombre d'indicateurs sélectionnés : 1649



Des termes évocateurs

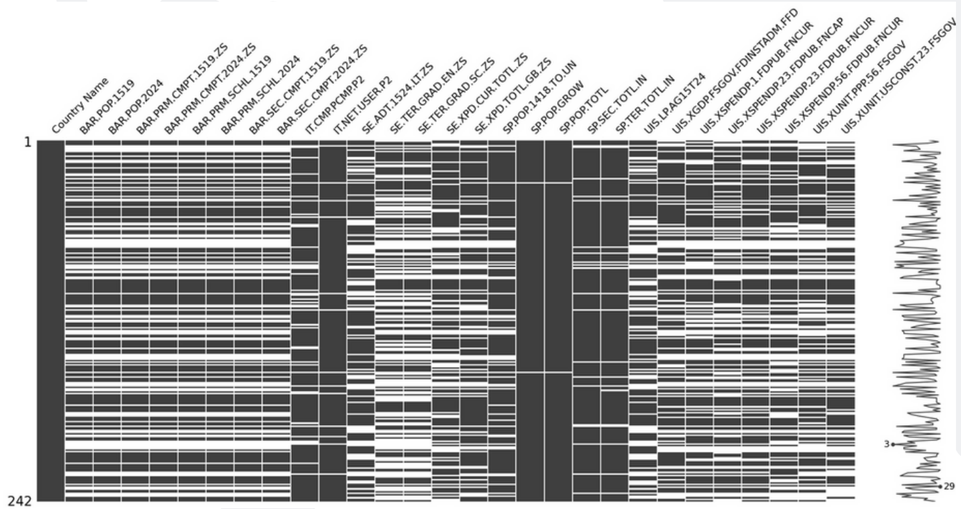
Dataframe de travail

```
Indicator Name
Population growth (annual %)
Population, total
Internet users (per 100 people)
Population of the official age for tertiary education, both sexes (number)
Population of the official age for secondary education, both sexes (number)
Personal computers (per 100 people)
Population, ages 14-18, total
Youth literacy rate, population 15-24 years, both sexes (%)
Expenditure on education as % of total government expenditure (%)
Youth illiterate population, 15-24 years, both sexes (number)
Government expenditure in educational institutions as % of GDP (%)
Current expenditure as % of total expenditure in public institutions (%)
Current expenditure as % of total expenditure in secondary public institutions (%)
Current expenditure as % of total expenditure in primary public institutions (%)
Current expenditure as % of total expenditure in tertiary public institutions (%)
Government expenditure per tertiary student (PPP$)
Government expenditure per secondary student (constant US$)
Capital expenditure as % of total expenditure in secondary public institutions (%)
Barro-Lee: Population in thousands, age 15-19, total
Barro-Lee: Population in thousands, age 20-24, total
Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary
Barro-Lee: Percentage of population age 20-24 with primary schooling. Completed Primary
Barro-Lee: Percentage of population age 15-19 with secondary schooling. Completed Secondary
Barro-Lee: Percentage of population age 15-19 with primary schooling. Completed Primary
Barro-Lee: Average years of primary schooling, age 20-24, total
Barro-Lee: Average years of primary schooling, age 15-19, total
Percentage of graduates from tertiary education graduating from Engineering, Manufacturing and Construction programmes, both sexes (%)
Percentage of graduates from tertiary education graduating from Science programmes, both sexes (%)
```

préselection de 29 indicateurs sur critères métiers

pivot sur clé ‘Country Name’

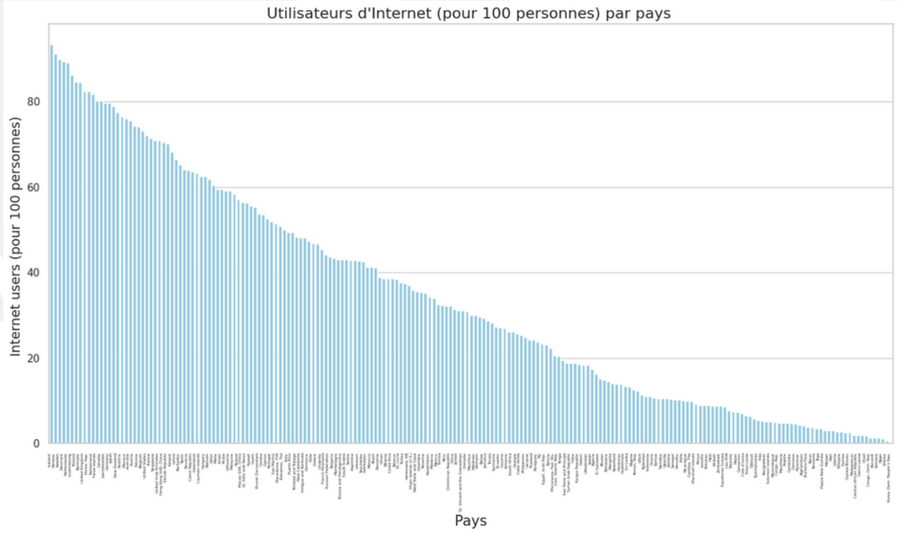
```
#création du nouveau DataFrame df_info
df_info = df_indicateurs.pivot(index='Country Name', columns='Indicator Name', values='Periode 2005 - 2015')
df_info_code = df_indicateurs.pivot(index='Country Name', columns='Indicator Code', values='Periode 2005 - 2015')

#réinitialisation de l'index pour avoir "Country Name" comme une colonne normale
df_info.reset_index(inplace=True)
df_info_code.reset_index(inplace=True)


#affichage du nouveau DataFrame
df_info_code
```

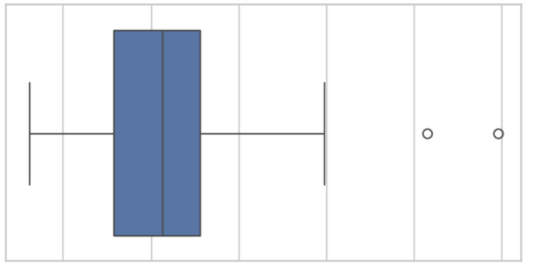


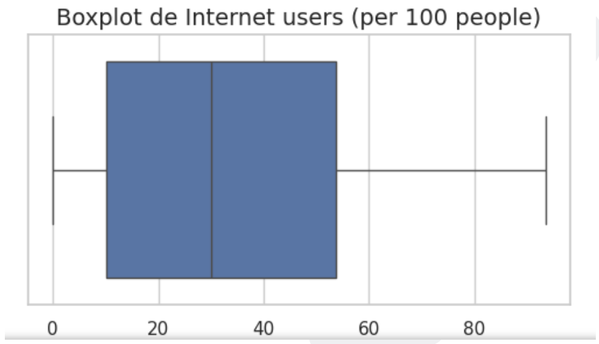
Création du df intermédiaire pour nettoyage

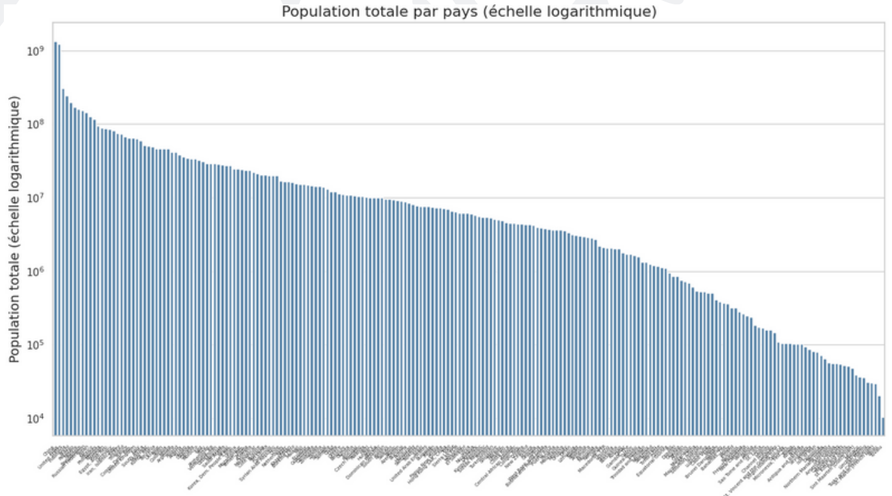
- outliers/valeurs aberrantes
- valeurs manquantes

Recherche outliers

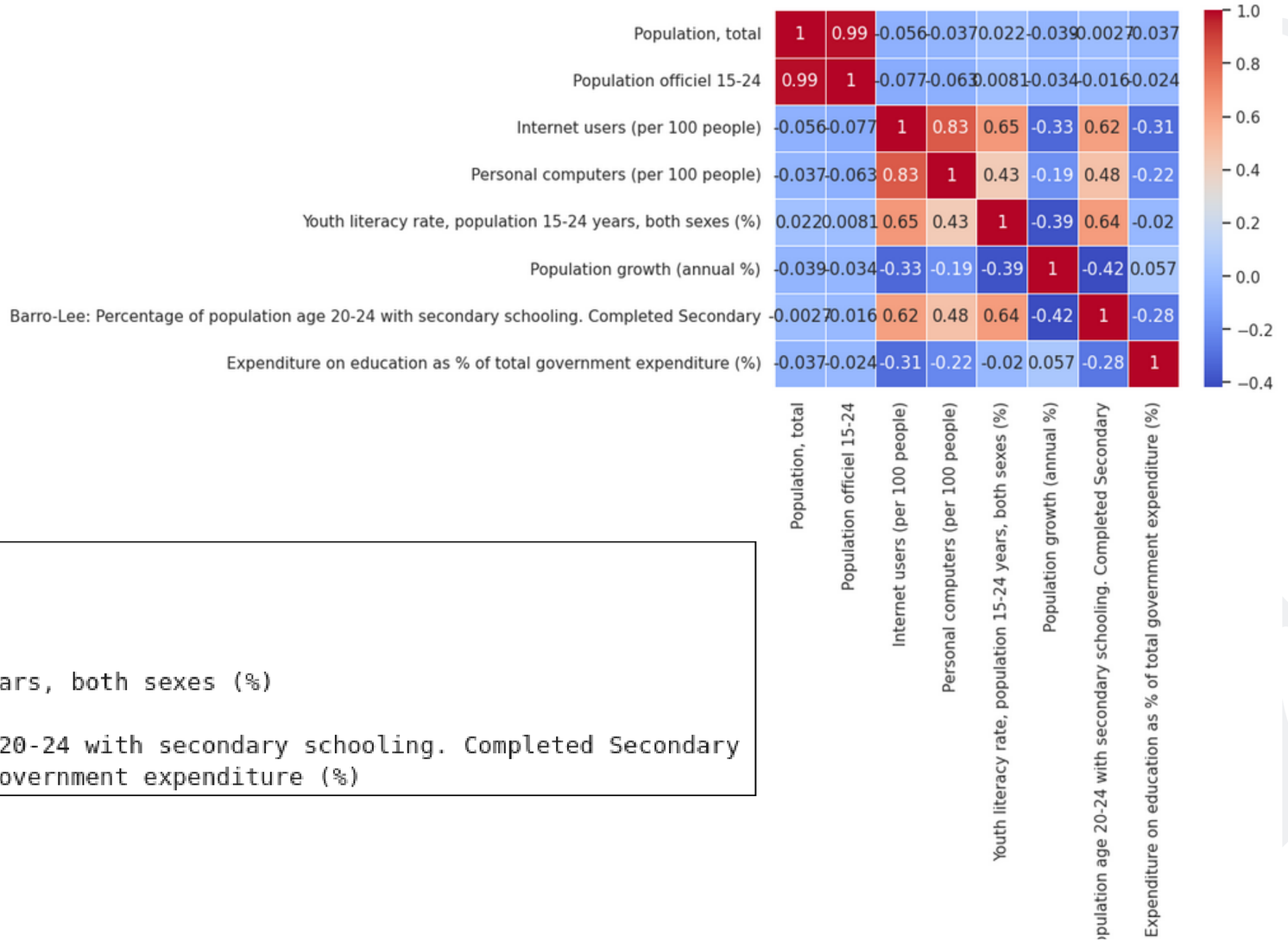








Affinage sélection indicateurs




```
Population, total
Population officiel 15-24
Internet users (per 100 people)
Personal computers (per 100 people)
Youth literacy rate, population 15-24 years, both sexes (%)
Population growth (annual %)
Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary
Expenditure on education as % of total government expenditure (%)
```

Traitement valeurs manquantes

```
# Remplacement des valeurs NaN par la moyenne par groupe identique selon critère Region et Income Group

# Liste des colonnes quantitatives avec des valeurs manquantes
colonnes_quantitatives = [
    'Population, total',
    'Population officiel 15-24',
    'Internet users (per 100 people)',
    'Personal computers (per 100 people)',
    'Youth literacy rate, population 15-24 years, both sexes (%)',
    'Population growth (annual %)',
    'Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary',
    'Expenditure on education as % of total government expenditure (%)'
]

# Remplacement des
for colonne in colo
    df_indics[color
    #impression du
    print(df_indics
```

```
#Remplacement valeur nan présente dans une colonne par la moyenne de la colonne afin d'obtenir un score synthétique sans impact

#extraction liste nom de colonnes
df_indics.columns.to_list()

#sélection des colonnes à traiter
colonnes_a_traiter = [
    'Population, total',
    'Population officiel 15-24',
    'Internet users (per 100 people)',
    'Personal computers (per 100 people)',
    'Youth literacy rate, population 15-24 years, both sexes (%)',
    'Population growth (annual %)',
    'Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary',
    'Expenditure on education as % of total government expenditure (%)'
]

#remplacement des valeurs NaN par la moyenne de chaque colonne
for colonne in colonnes_a_traiter:
    moyenne_colonne = df_indics[colonne].mean()
    df_indics[colonne].fillna(moyenne_colonne, inplace=True)

#affichage du dataframe mis à jour
df_indics
```

Scoring

Population, total

Population officiel 15-24

→ **Score_pop / 12 pts**

Internet users (per 100 people)

Personal computer (per 100 people)

→ **Score_tech / 4 pts**

Youth literacy rate, population 15-24 years, both sexes (%)

Barro-Lee: Percentage of population age 20-24 with secondary schooling. Completed Secondary

→ **Score_edu / 6 pts**

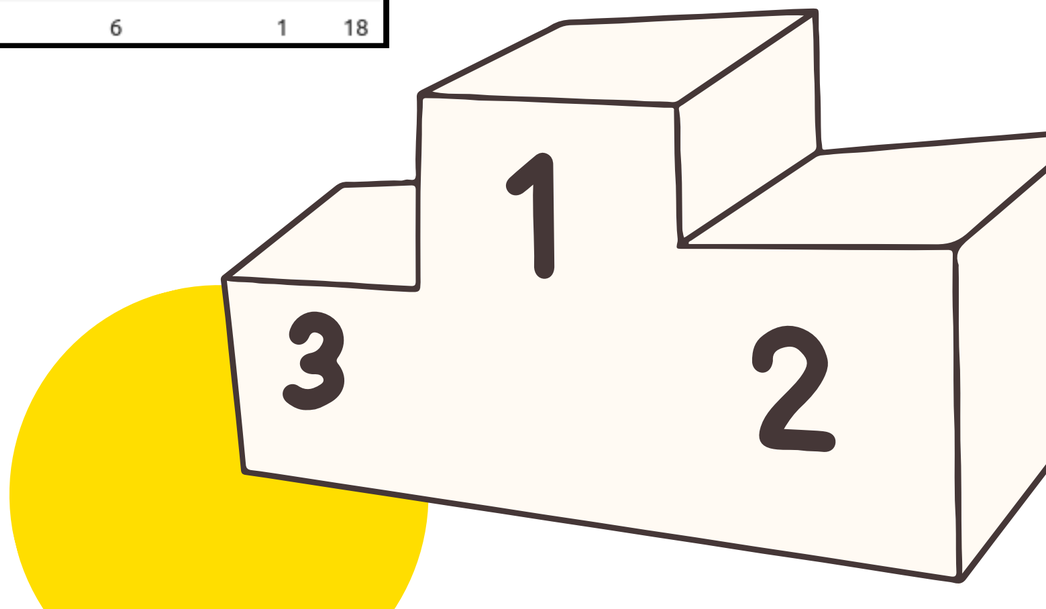
Expenditure on education as % of total government expenditure (%)

Population growth (annual %)

→ **Score_future / 3 pts**

Un classement représentatif

Country Name	Country Code	Region	Income Group	Population	Technologie	Education	Perspectives	Total
Australia	AUS	East Asia & Pacific	High income: OECD	8	4	6	2	20
Germany	DEU	Europe & Central Asia	High income: OECD	9	4	6	0	19
Saudi Arabia	SAU	Middle East & North Africa	High income: nonOECD	8	3	5	3	19
China	CHN	East Asia & Pacific	Upper middle income	12	2	4	1	19
United Kingdom	GBR	Europe & Central Asia	High income: OECD	9	4	5	1	19
South Africa	ZAF	Sub-Saharan Africa	Upper middle income	9	2	6	2	19
United States	USA	North America	High income: OECD	11	4	3	1	19
Singapore	SGP	East Asia & Pacific	High income: nonOECD	7	4	5	3	19
Belgium	BEL	Europe & Central Asia	High income: OECD	7	4	6	1	18

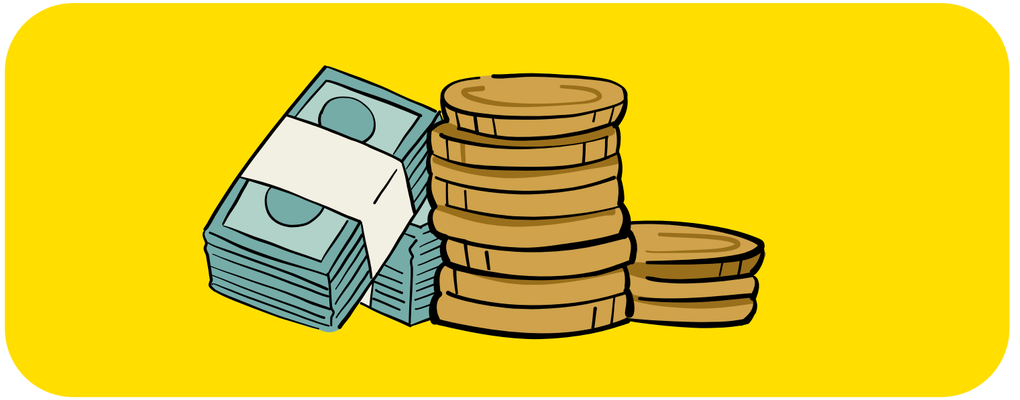


Des critères supplémentaires

Afin de rendre l'analyse et la discussion dynamique, on va intégrer le filtrage par Région et par Groupe de Revenu

Région

Richesse

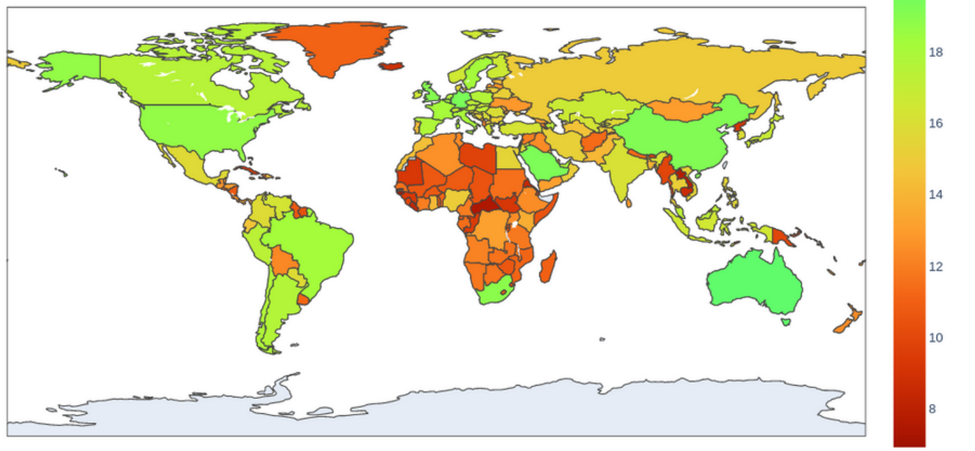


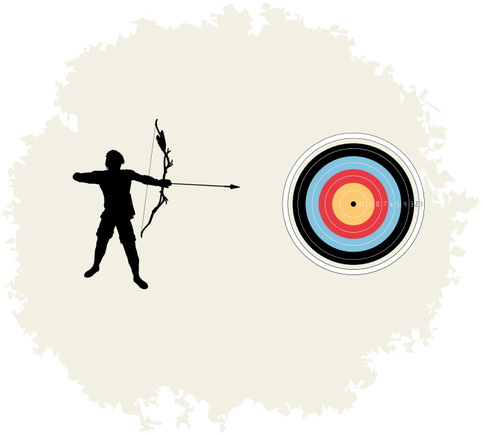
Un outil sur mesure

Monde entier

Toutes les catégories

Carte Choroplèthe - Scores Totals par Pays (Tous, Toutes)

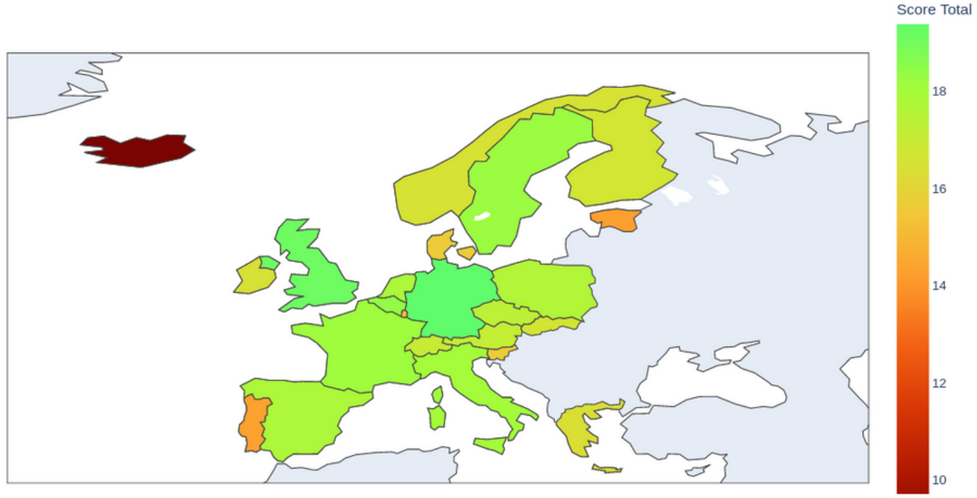




Europe & Central Asia

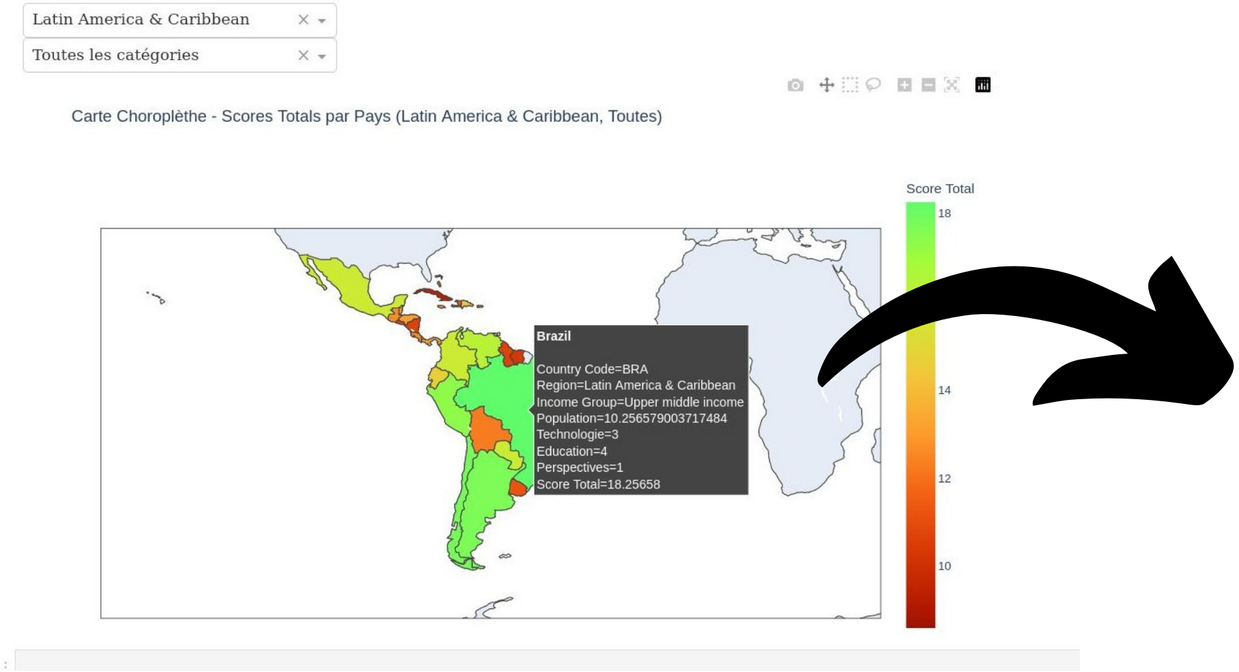
High income: OECD

Carte Choroplèthe - Scores Totals par Pays (Europe & Central Asia, High income: OECD)



ciblage dynamique

Un outil sur mesure



Des pistes d’amélioration...

- Prendre en compte les dynamiques évolutives des valeurs
- Pondérer les scores en fonction des évolutions
- Améliorer l’affichage des scores en permettant l’affichage selon une ou plusieurs catégories
- Permettre à l’utilisateur de pondérer lui même les scores

Merci pour votre **Attention**