SOUTENANCE
PROJET N°2

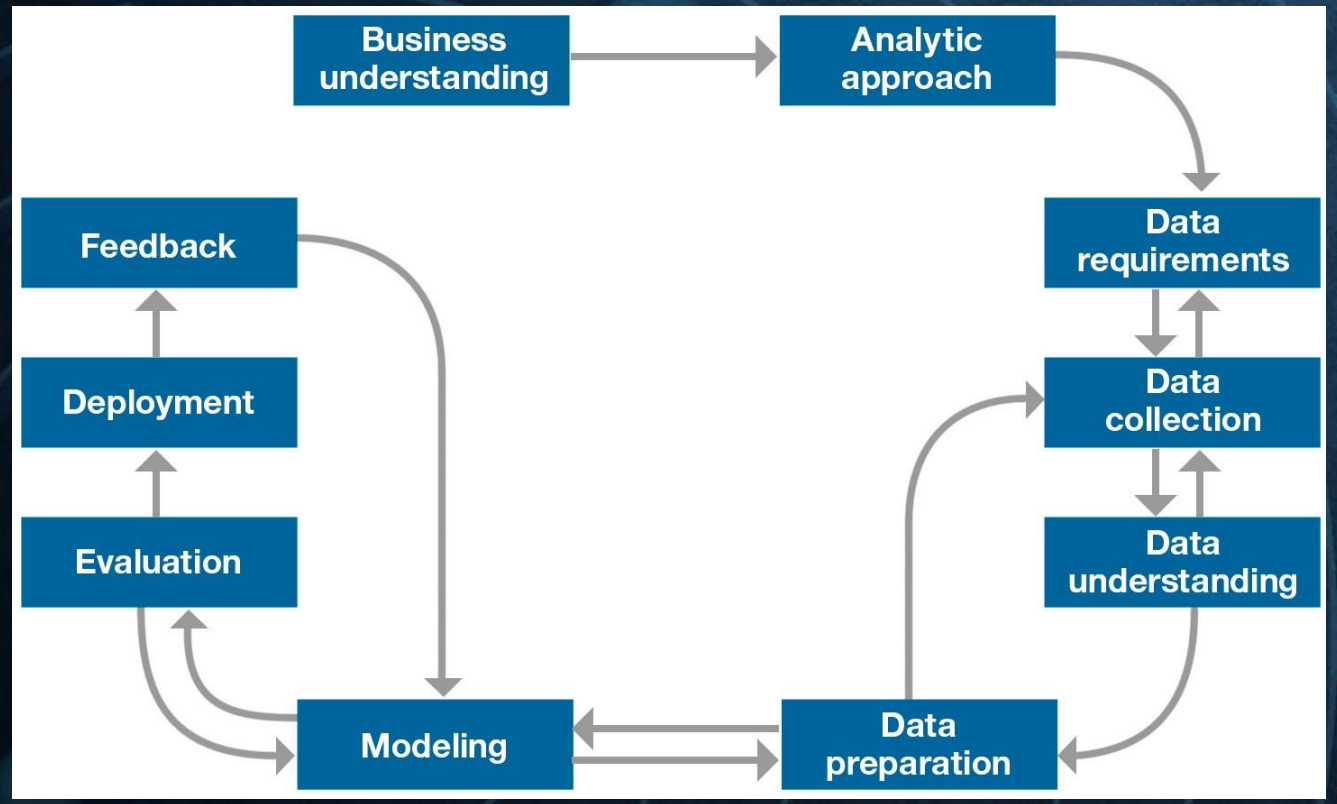
« ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS»

Florian Judée

02 Février 2021

OPENCLASSROOMS

PROGRAMME



"Like traditional scientists, data scientists need a foundational methodology that will serve as a guiding strategy for solving problems"

John Rollins
Data Scientist, IBM Analytics, IBM

ETAPE 1 : COMPRÉHENSION COMMERCIALE

LE CLIENT :

- ACADEMY est une start-up de la EdTech
- Il fournit du contenu de formation en France
- Niveau lycée et université
- Volonté d'expansion à l'international

PROBLÉMATIQUE DU CLIENT:

« Dans un contexte d'expansion à l'international, l'entreprise cherche a déterminer une liste de pays avec un fort potentiel de clients .

Le client cherche également à connaitre l'évolution de ce potentiel à moyen et long terme.»

ETAPE 2 : APPROCHE ANALYTIQUE

LE DATASET :

- Le dataset étant fournit par notre client nous devrons d'abord le comprendre, le nettoyer et définir si il est suffisamment complet pour réaliser l'étude.

ETUDE PRÉ-EXPLORATOIRE:

- Recherche d'éléments discriminants pour l'objectif du client,
- Mise en forme du dataset original (réduction aux données essentielles) contenant une liste de pays initiale
- Elimination des pays à faible attractivité par études successives des facteurs d'impacts
- Déterminer un moyen de hiérarchiser l'attractivité des pays répondant à tous les critères
- Etude de l'évolution temporelle à moyen et long termes du potentiel de client

LIVRABLES:

- Liste de pays hiérarchisé en fonction de leurs attractivités

ETAPE 3 : BESOIN DE DONNÉES

- Liste de pays conséquente
- Information sur les potentiels clients :
  - Nombre de personnes en âge d'accéder au contenu
  - Evolution de la population cible
  - Disposent ils d'un accès au numérique (ordinateur et internet)
  - Ont-ils les moyens financier d'accéder au service du client
- Information sur les pays:
  - Si le pays dispose d'une infrastructure scolaire

SYNTHÉSE DES BESOINS:

« Dans le cadre de notre étude nous avons besoin d'avoir accès à des données de types éducatives, sociales et économique.
Ces données doivent être référencées temporellement et géographiquement»

ETAPE 4 : COLLECTE DE DONNÉES

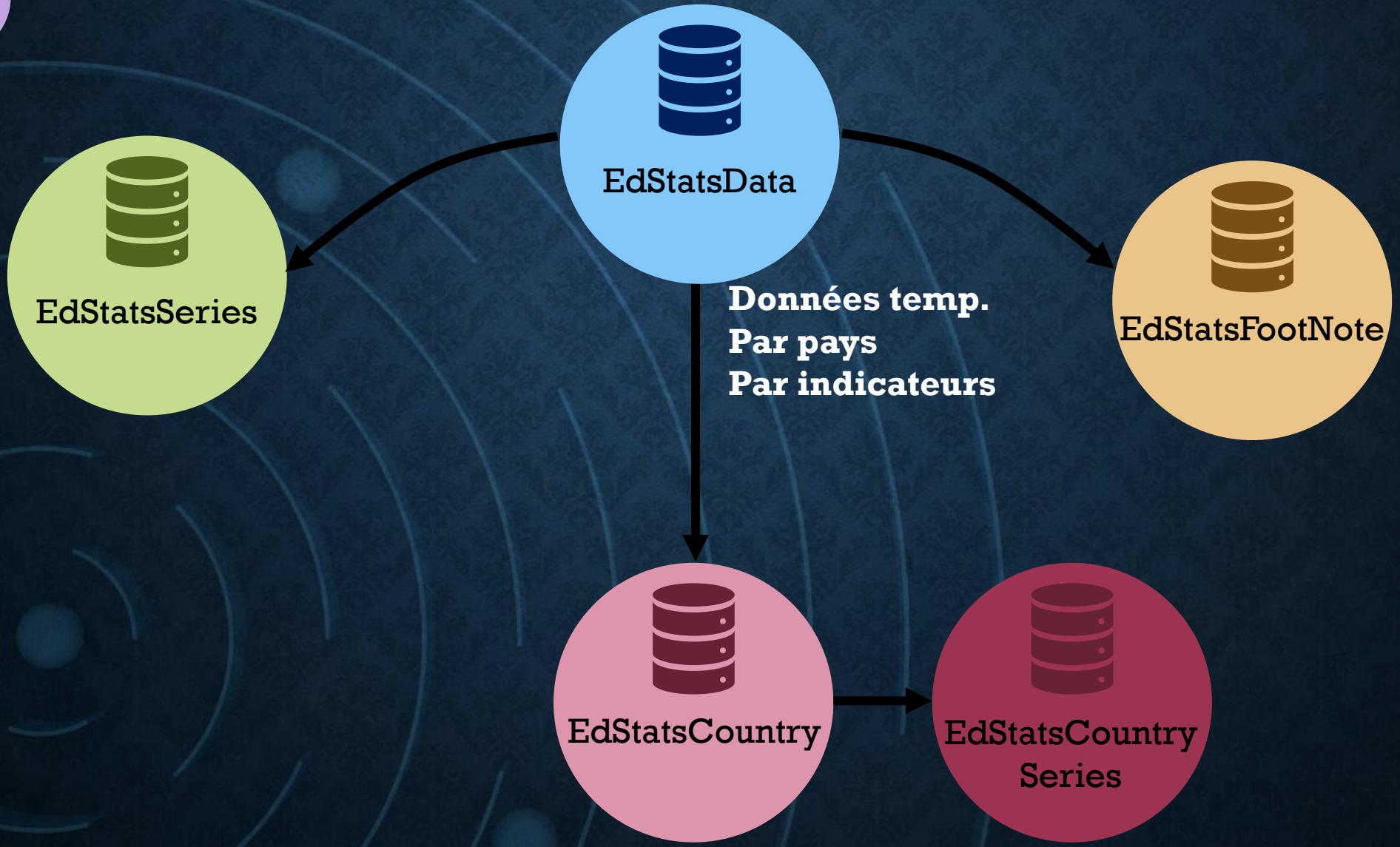
ORIGINE DU DATASET :

- **La Banque Mondiale** est un groupe regroupant 189 pays
- Il fournit un financement aux projets de développement
- Donne accès librement à des statistiques dans le monde

DESCRIPTIF DU DATASET:

« Parmi toutes les données dont il dispose, le site met à dispositif des statistiques sur l'éducation à travers le monde via 4000 indicateurs différents.
Ces indicateurs couvrent le cycle d'éducation pré-primaire au supérieur.
Le dataset EdStats, fournit par notre client, donne également des informations passées ainsi que des projections.»

ETAPE 5 : COMPRÉHENSION DES DONNÉES



ETAPE 5 : COMPRÉHENSION DES DONNÉES

EdStatsData	**Données numériques sur l'évolution temporelle d'indicateurs par pays** Taille : 886 930 lignes, 70 colonnes Nombreuses données manquantes (86%), aucun doublon
EdStatsSeries	**Informations sur les indicateurs présents dans EdStatsData** Taille : 3665 lignes, 21 colonnes Nombreuses données manquantes (70%), aucun doublon
EdStatsFootNote	**Informations sur les données numériques présentes dans EdStatsData** Taille : 643 638 lignes, 5 colonnes Nombreuses données manquantes (20%, 1 colonne), aucun doublon
EdStatsCountry	**Informations sur les pays présents dans EdStatsData** Taille : 241 lignes, 32 colonnes Nombreuses données manquantes (50%), aucun doublon
EdStatsCountrySeries	**Informations sur la source des données de EdStatsCountry** Taille : 613 lignes, 4 colonnes Nombreuses données manquantes (25% 1 colonne), aucun doublon

ETAPE 5 : COMPRÉHENSION DES DONNÉES

EdStatsData

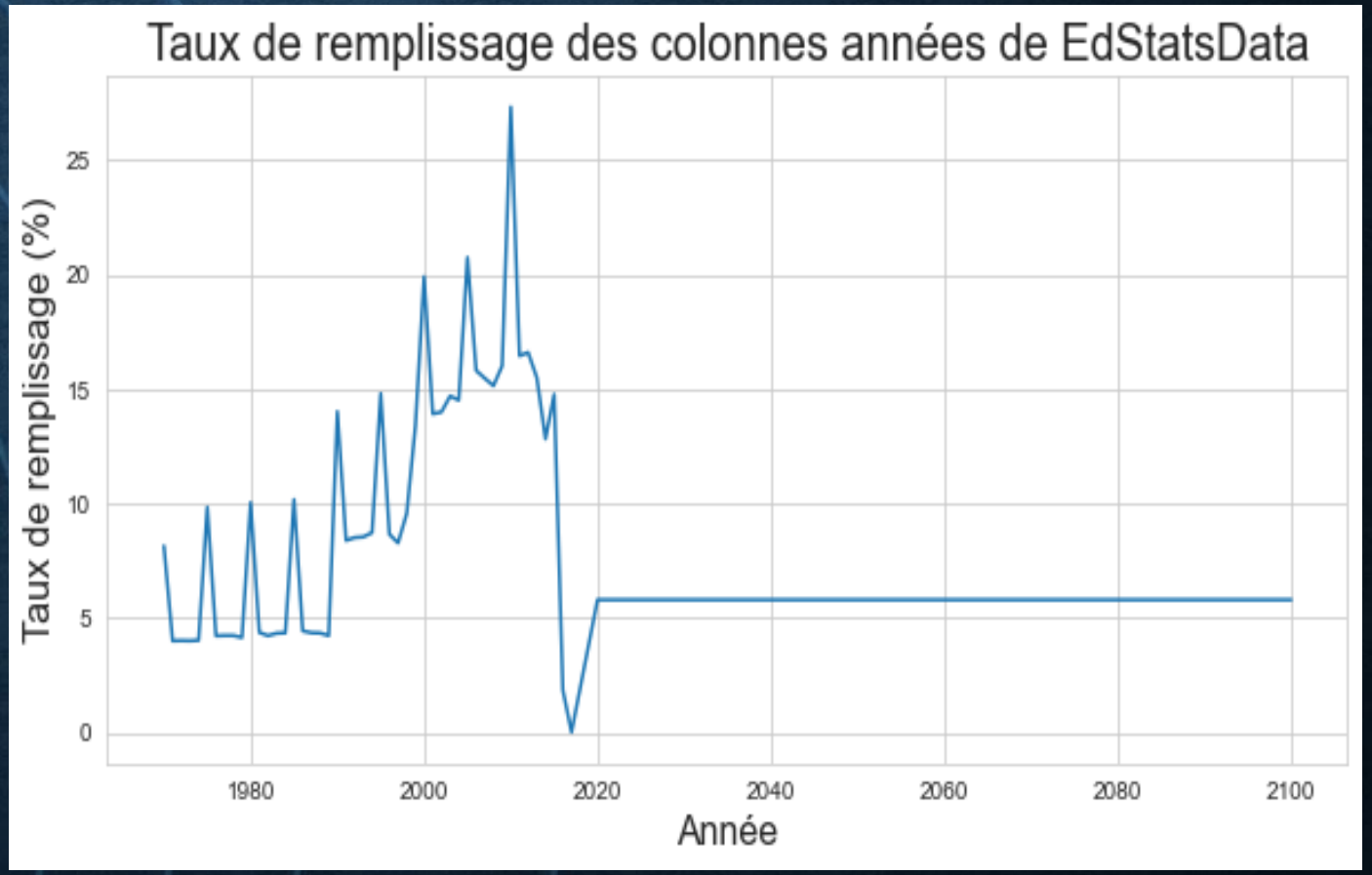
241 pays différents

3665 indicateurs différents

Données de 1970 à 2017

Prévisions de 2020 à 2100 (5 ans)

Nan : 86 %



ETAPE 5 : COMPRÉHENSION DES DONNÉES

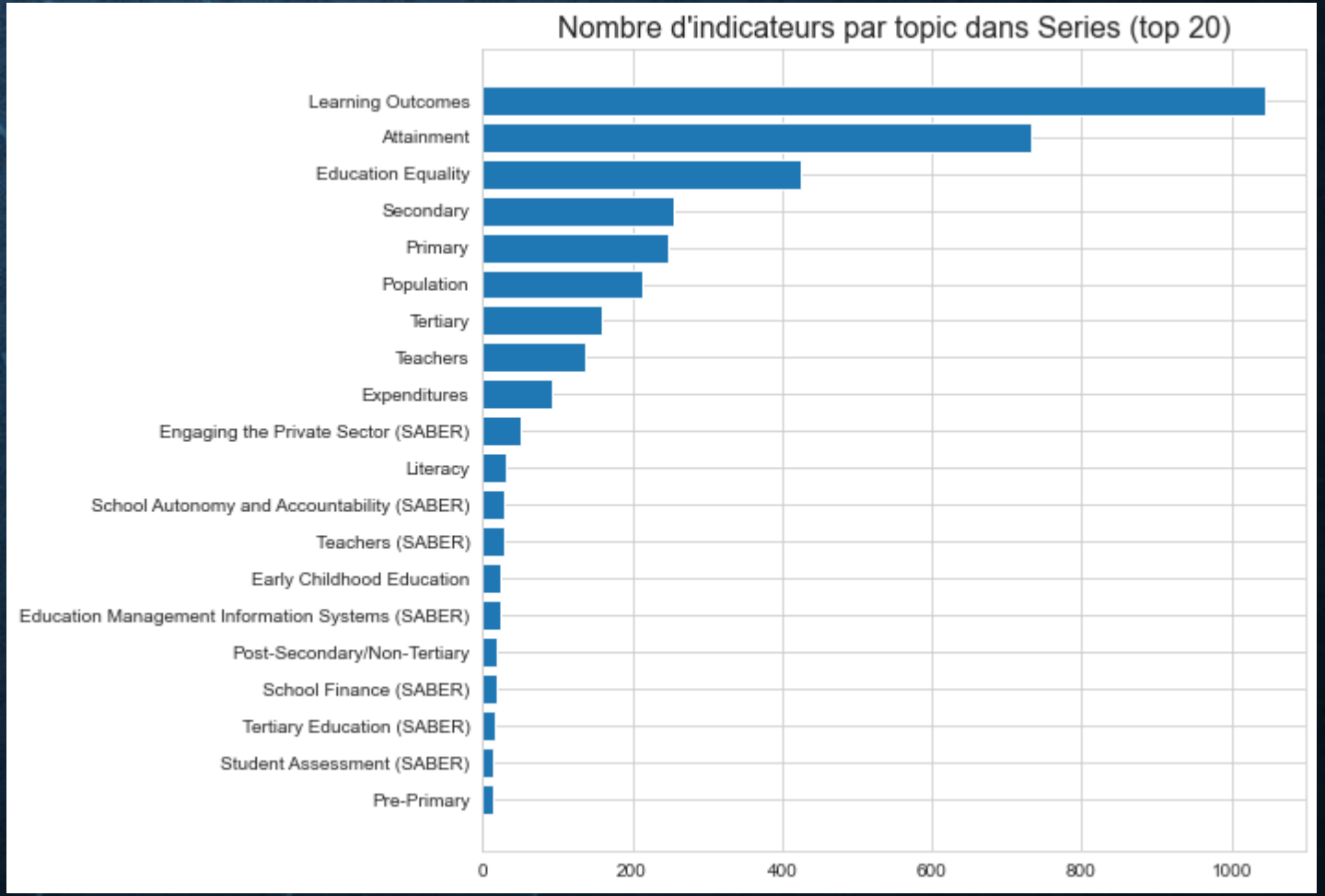
EdStatsSeries

3665 indicateurs différents

37 Topics (éducation, santé …)

Nan : 70 %

Pas de données manquantes dans les colonnes importantes



ETAPE 5 : COMPRÉHENSION DES DONNÉES

EdStatsCountry

241 pays différents

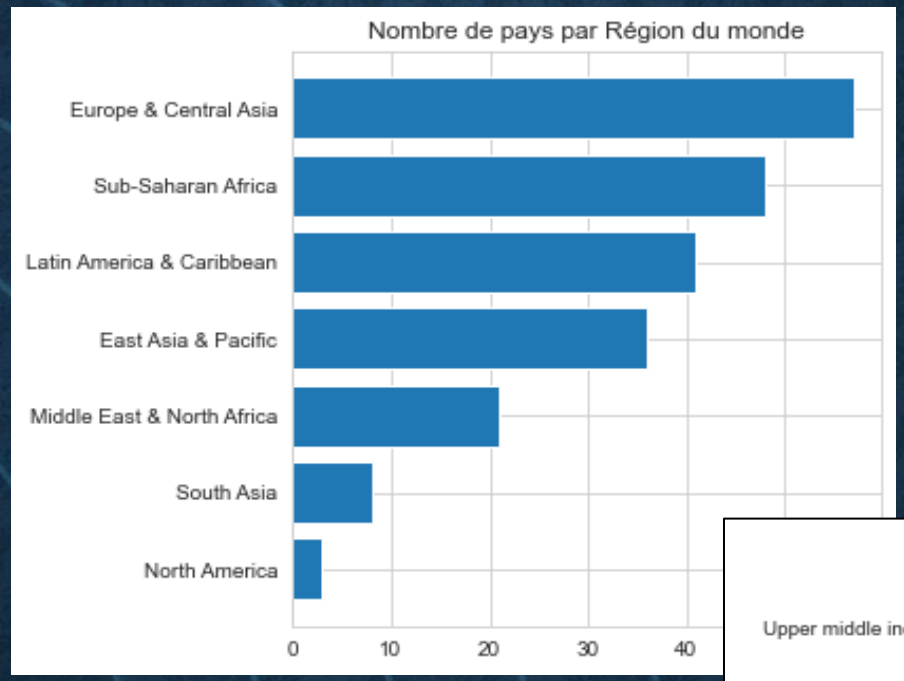
Dont 27 groupements

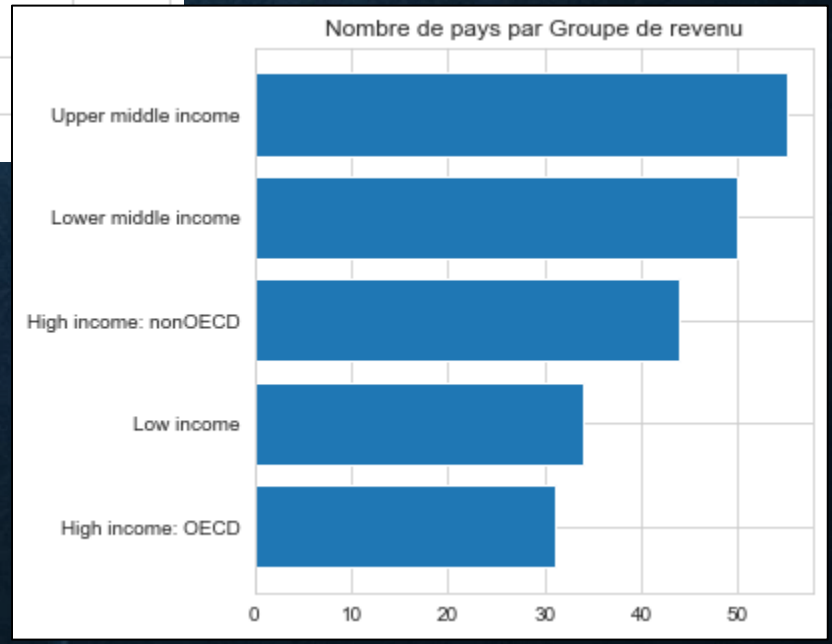
7 régions du monde

5 groupes de revenus

Nan : 50 %

Pas de données manquantes dans les colonnes importantes





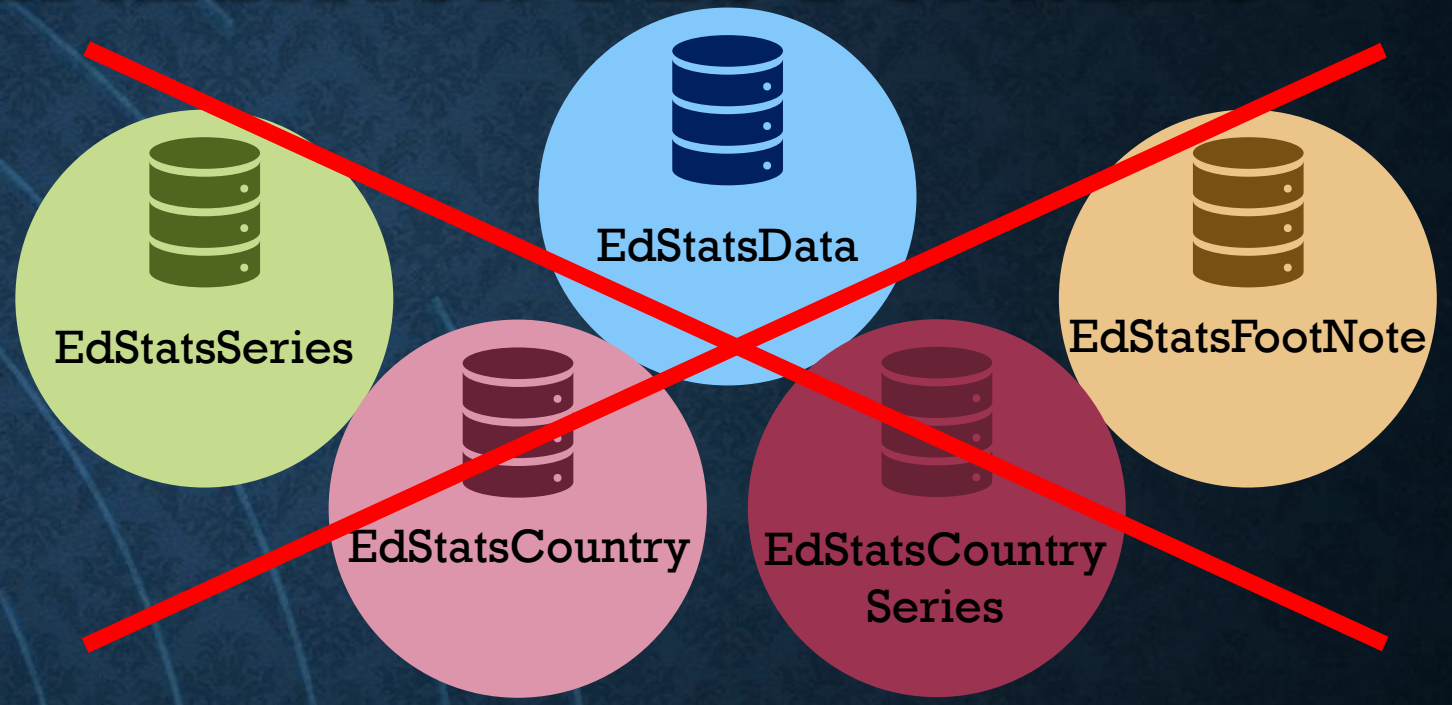
ETAPE 6 : PRÉPARATION DES DONNÉES

OPERATIONS:

1. Supprimer les datasets inutiles
2. Supprimer les colonnes inutiles (NaN)
3. Faire une jointure des colonnes essentielles
4. Regrouper les colonnes années par décennies

RESULTATS:

« Nous avons maintenant un seul fichier cleané qui nous sert pour les prochaines étapes»



Taille : 886 930 lignes
20 colonnes (-60)
157 Mo (43%)

ETAPE 7 : EXTRACTION DES DONNÉES

OPERATIONS:

A partir de l'étape 3 (Besoin de données), nous avons identifier des critères de sélections pour notre projet.

Recherche de ces critères dans la base de donnée.

1. **IT.Net.User.P2** Taux d'internet pour 100 personnes
2. **IT.CMP.PCMP.P2** Nombre d'ordinateur personnel pour 100 personnes
3. **NY.GDP.MKTP.PP.CD** PIB ppp
4. **SP.POP.1524.TO.UN** total de la population agée de 15 à 24 ans
5. **SP.POP.1015.TO.UN** total de la population agée de 10 à 15 ans
6. **SP.POP.GROW** augmentation de la population…
7. **SP.POP.TOTL** total de la population
8. **SE.TER.ENRL** Personnes inscrites à l'université
9. **UIS.E.3** Personnes inscrites au lycée
10. **UIS.E.4** Personnes inscrites en formation post-bac

ETAPE 7 : EXTRACTION DES DONNÉES

Combien de nos pays (sur 242 présents)

ont une donnée pour chacun de nos indicateurs ?

	Indicator Name	Indicator Code	2010s
0	Population growth (annual %)	SP.POP.GROW	240
1	Population, total	SP.POP.TOTL	240
2	Internet users (per 100 people)	IT.NET.USER.P2	229
3	GDP, PPP (current international $)	NY.GDP.MKTP.PP.CD	217
4	Enrolment in upper secondary education, both s...	UIS.E.3	206
5	Enrolment in tertiary education, all programme...	SE.TER.ENRL	197
6	Population, ages 10-15, total	SP.POP.1015.TO.UN	181
7	Population, ages 15-24, total	SP.POP.1524.TO.UN	181
8	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	137
9	Personal computers (per 100 people)	IT.CMP.PCMP.P2	0

	Indicator Name	Indicator Code	2050s
0	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	0
1	Enrolment in tertiary education, all programme...	SE.TER.ENRL	0
2	Enrolment in upper secondary education, both s...	UIS.E.3	0
3	GDP, PPP (current international $)	NY.GDP.MKTP.PP.CD	0
4	Internet users (per 100 people)	IT.NET.USER.P2	0
5	Personal computers (per 100 people)	IT.CMP.PCMP.P2	0
6	Population growth (annual %)	SP.POP.GROW	0
7	Population, ages 10-15, total	SP.POP.1015.TO.UN	0
8	Population, ages 15-24, total	SP.POP.1524.TO.UN	0
9	Population, total	SP.POP.TOTL	0

BILAN:

« Il n'existe aucune valeur pour l'indicateur d'accès à l'ordinateur => à supprimer
Aucun de nos indicateurs n'a de prédiction après les années 2010s => autre solution
Nombre de pays par indicateurs fluctuant => supprimer les pays sans données »

ETAPE 7 : EXTRACTION DES DONNÉES

HYPOTHESES:

Quels sont les indicateurs essentielles ?

1. Supprimer les pays dont on à aucune information sur la population en lycée

2. Supprimer les pays sans valeurs d'accès à internet

	Indicator Name	Indicator Code	2010s
0	Enrolment in upper secondary education, both s...	UIS.E.3	201
1	Internet users (per 100 people)	IT.NET.USER.P2	201
2	Population growth (annual %)	SP.POP.GROW	201
3	Population, total	SP.POP.TOTL	201
4	GDP, PPP (current international $)	NY.GDP.MKTP.PP.CD	195
5	Enrolment in tertiary education, all programme...	SE.TER.ENRL	183
6	Population, ages 10-15, total	SP.POP.1015.TO.UN	162
7	Population, ages 15-24, total	SP.POP.1524.TO.UN	162
8	Enrolment in post-secondary non-tertiary educa...	UIS.E.4	130

RESULTATS:

« Nous sommes passé de 242 pays référencés à 201 (- 41) que nous pouvons utiliser »

ETAPE 7 : EXTRACTION DES DONNÉES

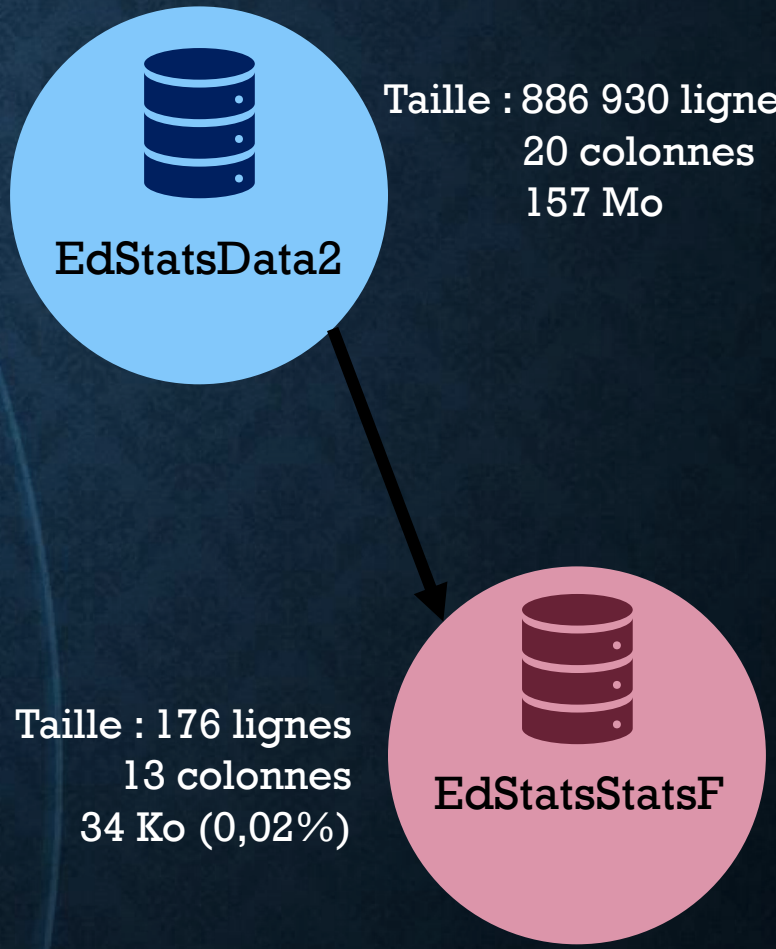
```
def mise_en_colonnes (df_lignes,annees='2010s'):
    for pays in Data_joined_corrected['Country Name'].unique():
        df_temp=Data_joined_corrected[(Data_joined_....
```

OPERATIONS:

1. Garder que les informations 2010s
2. Supprimer les groupements de pays
3. Mise en forme du dataframe
   1 pays par ligne
   1 indicateurs par colonne

RESULTATS:

« Nous disposons maintenant d'un fichier unique pour l'analyse de donnée qui est adapté à notre étude et beaucoup moins lourd»



ETAPE 8 : ANALYSER LES DONNÉES

```
def stats_describe (df):
    value=[]
    value.append((100-(df.isna().sum()/df.shape[0])*100).values[4:])
    ...
```

	Taux internet	PIB	Pop 1015	Pop 1524	Pop totale	Pop croissance	Evol pop 1024	Pop etudiants	Taux etudiants
Taux valeur %	100	96.5909	92.0455	92.0455	100	100	92.0455	54.5455	51.1364
Moyenne	44.3303	5.99484e+11	4.27617e+06	7.13549e+06	3.90416e+07	1.35466	8.49046	3.53774e+06	48.9078
Std	28.2743	1.98777e+12	1.49233e+07	2.56599e+07	1.44728e+08	1.33116	16.7152	1.21087e+07	25.1487
Minimum	0.894407	3.62263e+07	8817.5	13861.8	10815.6	-1.74492	-34.4941	2192	2.74209
Pays min	Eritrea	Tuvalu	Aruba	Aruba	Tuvalu	Syrian Arab Republic	Macao SAR, China	Andorra	Niger
Mediane	44.5875	6.96792e+10	832964	1.32233e+06	7.83316e+06	1.22867	6.5894	506981	52.5896
Maximum	96.5095	1.6886e+13	1.48501e+08	2.40163e+08	1.35772e+09	6.8515	50.4721	8.10319e+07	94.4762
Pays max	Iceland	China	India	India	China	Qatar	Niger	India	Finland

RESULTATS:

« La fonction crée permet de fournir des statistiques sur chacun des indicateurs sélectionnés. Elle permet également de vérifier si il existe des valeurs aberrantes. »c

ETAPE 8 : ANALYSER LES DONNÉES

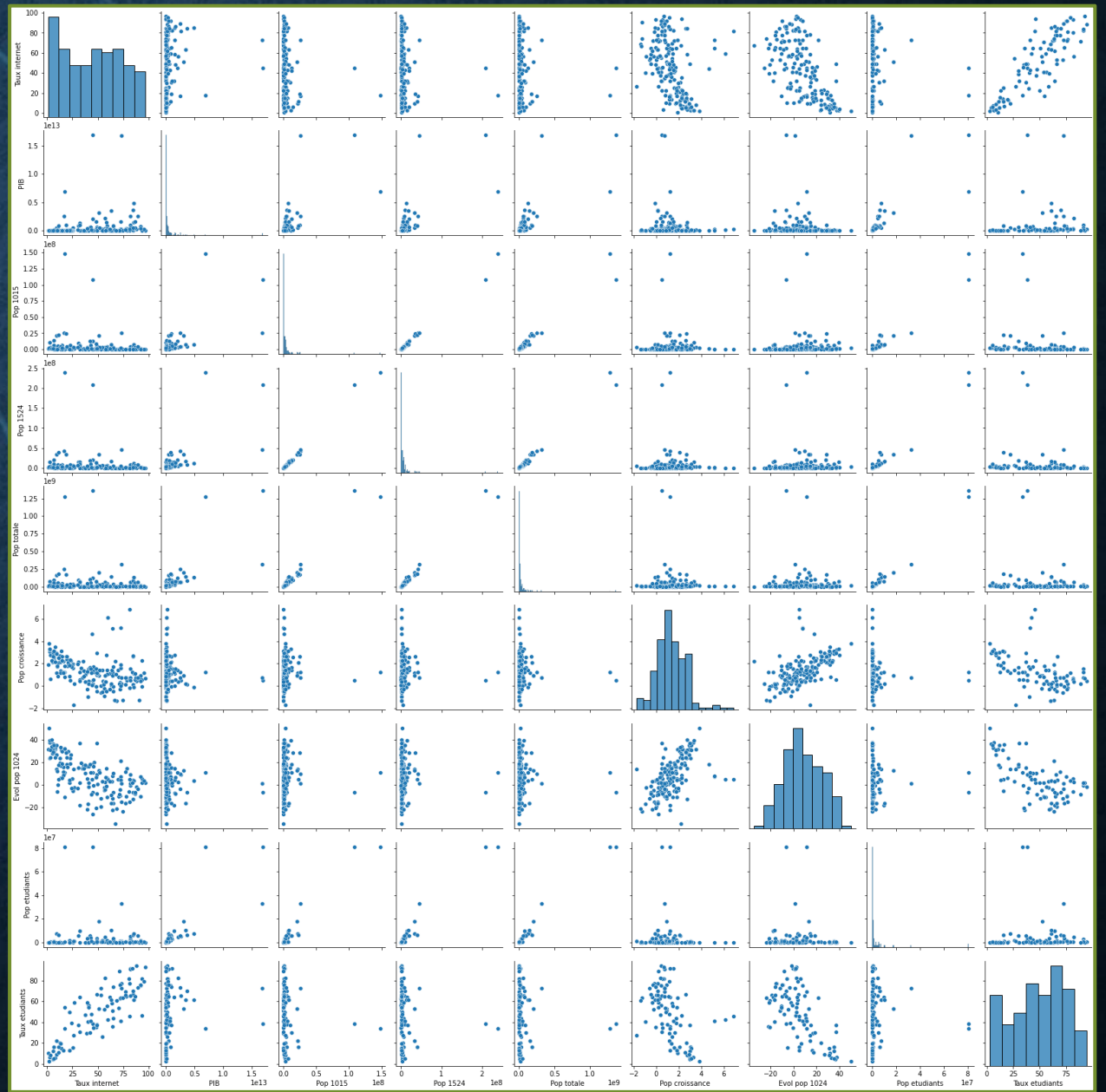
FONCTION PAIRPLOT:

1. Sur la diagonale nous obtenons les distribution de chacun de nos indicateur

2. Dans la matrice inferieure et supérieure, nous avons l'analyse bivariée de deux indicateurs.

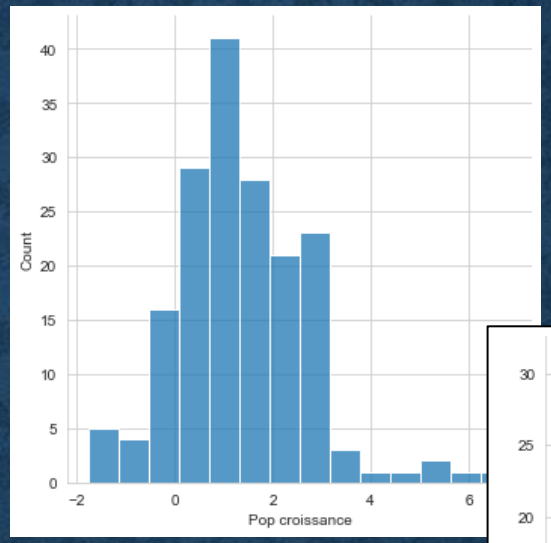
RESULTATS:

« Cela nous permets d'avoir une première estimation des corrélations entre indicateurs. Cela nous donne aussi une idée de la distribution de nos variables. »

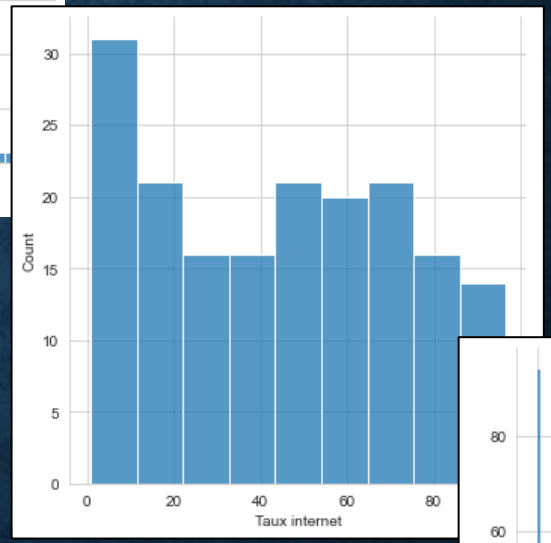


ETAPE 8 : ANALYSER LES DONNÉES

Croissance de la population (%)
SKEW $\gamma 1$ = 0,77
KURTOSIS $\gamma 2$ = 2,03



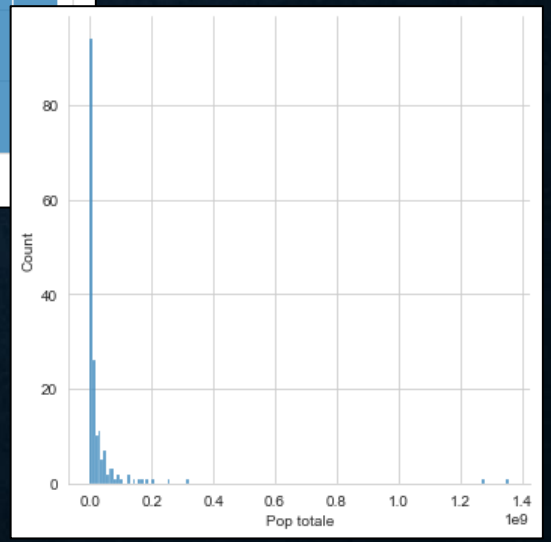
Taux accès internet (%)
SKEW $\gamma 1$ = 0,11
KURTOSIS $\gamma 2$ = -1,25



RESULTATS:

« L'analyse univarié de nos indicateurs montre trois distributions différentes.
Tout ce qui touche à la population devra être affiché en échelle logarithmique»

Population totale
SKEW $\gamma 1$ = 8
KURTOSIS $\gamma 2$ = 69



ETAPE 8 : ANALYSER LES DONNÉES

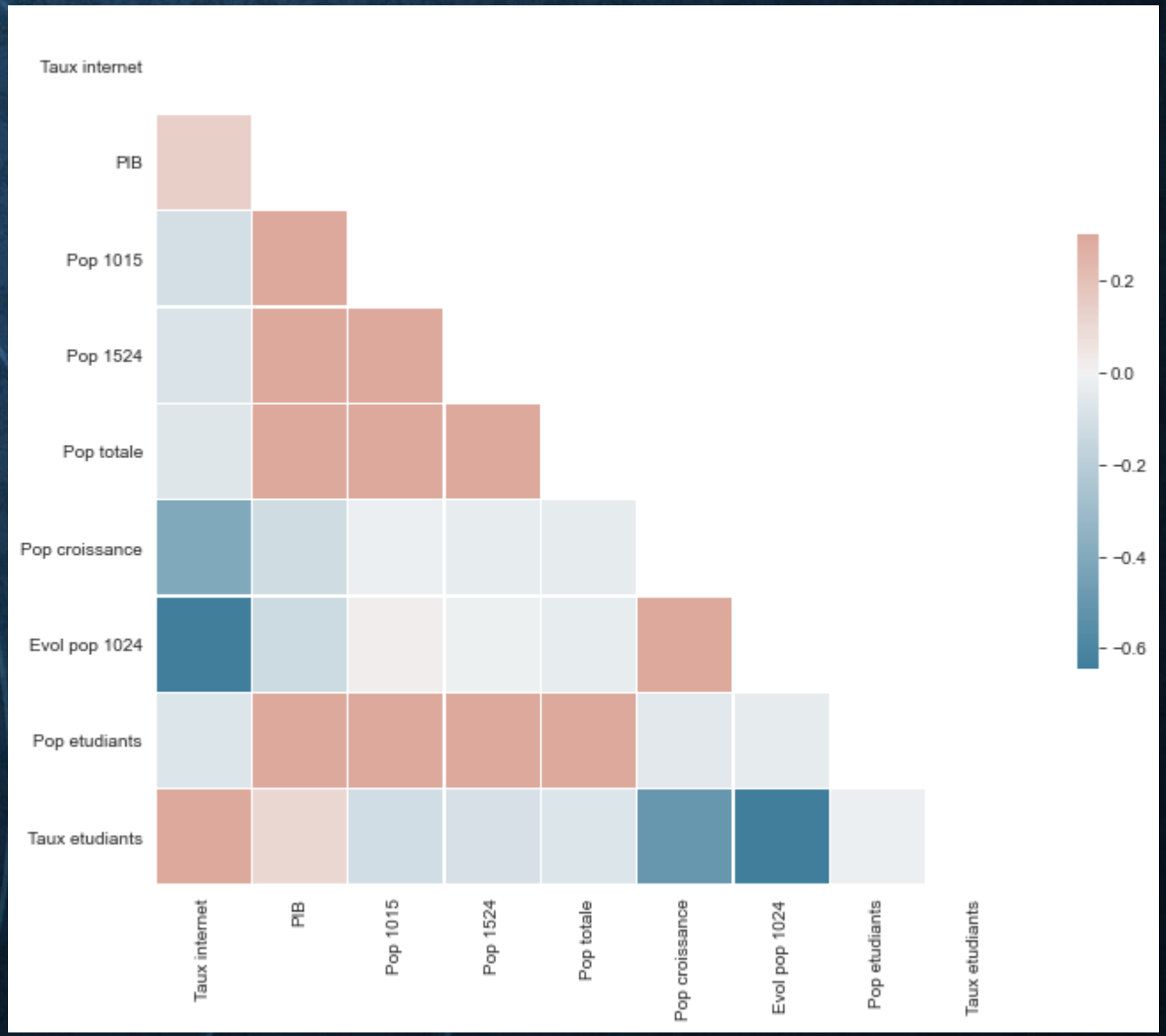
FONCTION HEATMAP:

Affiche en couleur l'indicateur de corrélation de Pearson r (-1 , 1) sur une matrice inférieure.

RESULTATS:

« Combiné à PAIRPLOT :
** Toute les Populations sont corrélées entre elles,*
** les deux indicateurs de croissance aussi,*
** corrélation entre Tx internet et étudiants avec la croissance de la population.*

Sert à définir mes prochains graphiques»



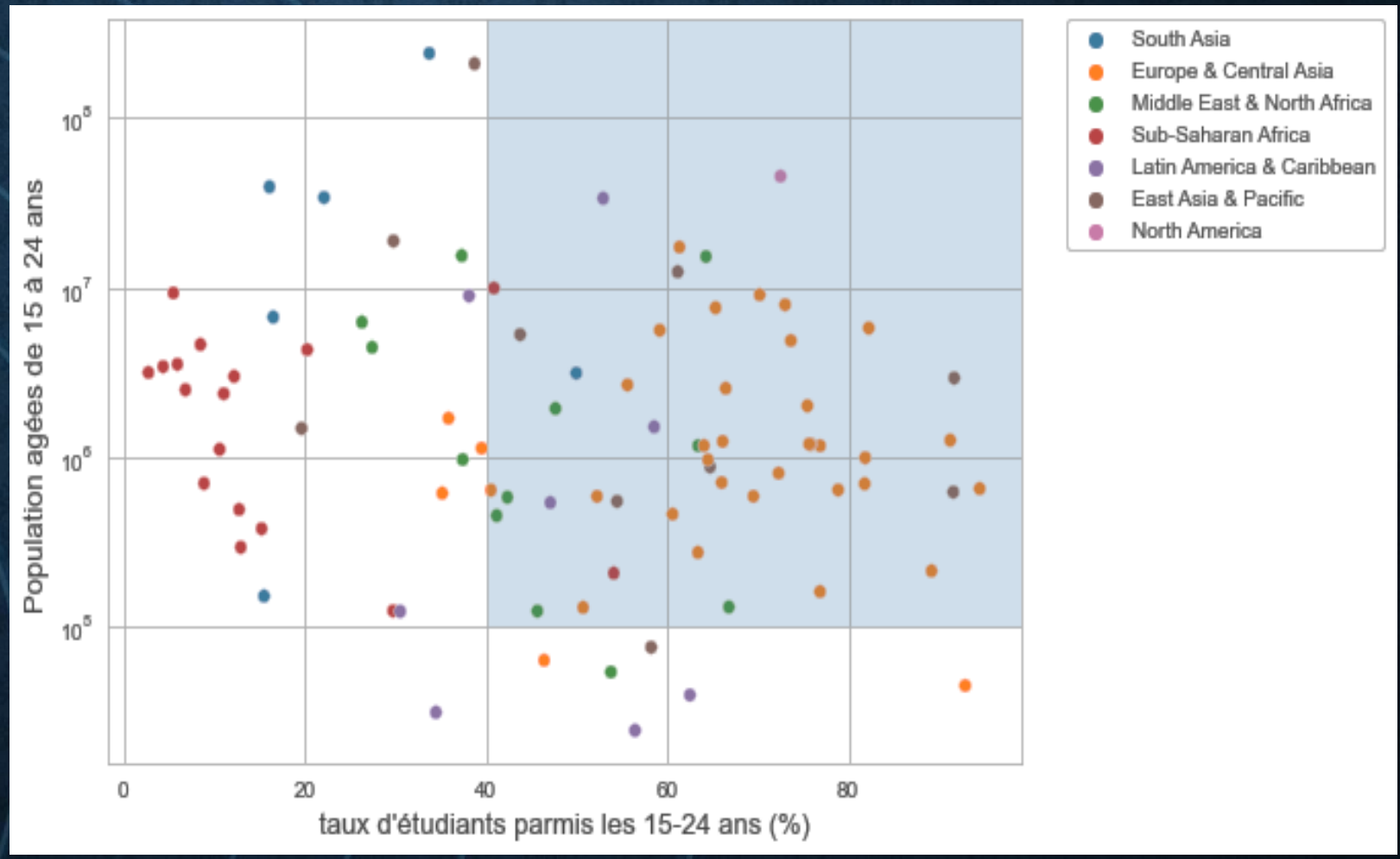
ETAPE 8 : ANALYSER LES DONNÉES

TRI N°1:

Population cible (15-24 ans) vs taux d'étudiants

RESULTATS:

« On passe de 176 pays référencés à seulement 51 dans notre liste de candidats potentiels (-125) »



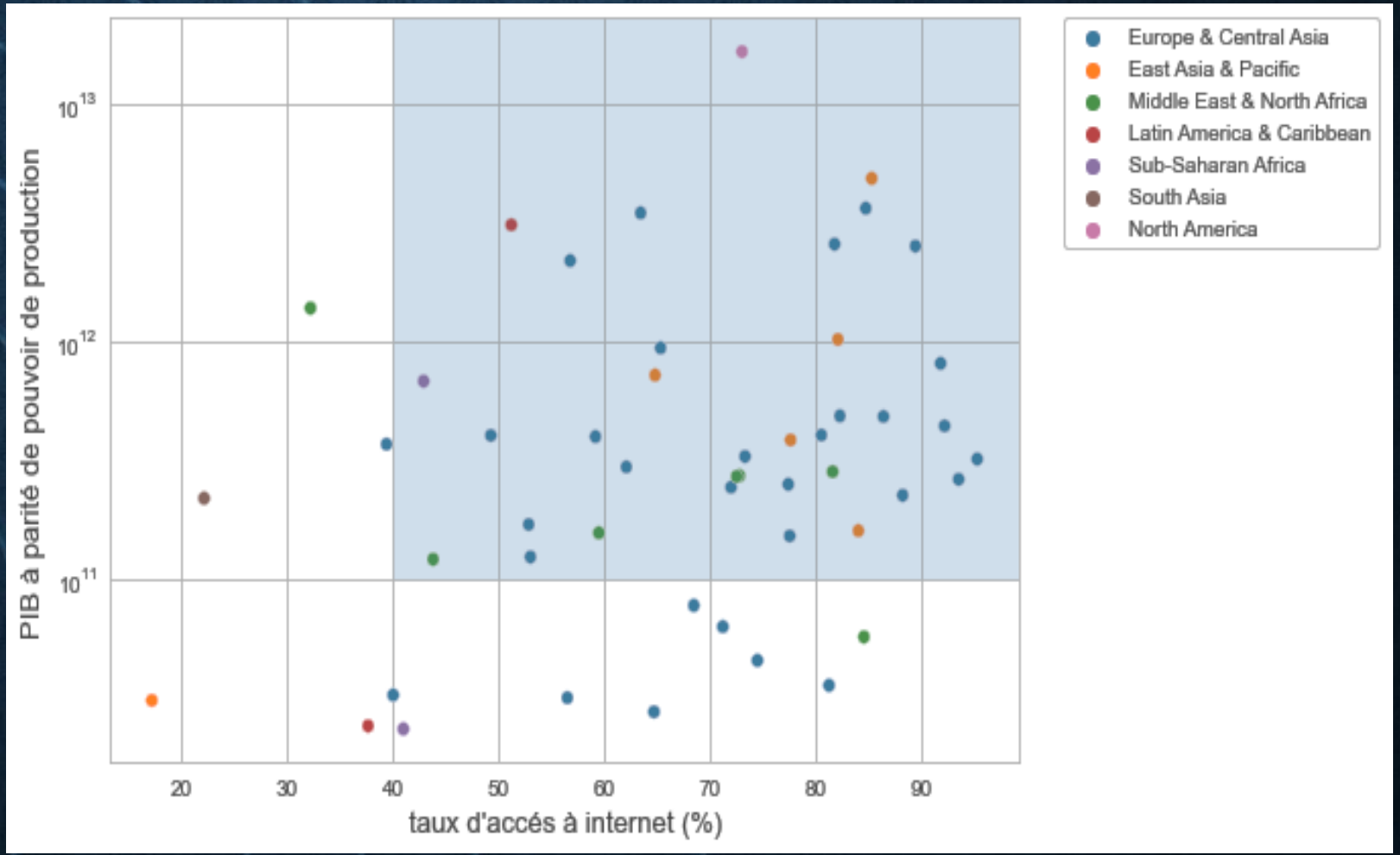
ETAPE 8 : ANALYSER LES DONNÉES

TRI N°2:

PIB (financier) vs taux d'accès à internet

RESULTATS:

« On passe d'une liste de 51 candidats à 36 pays (-15) »



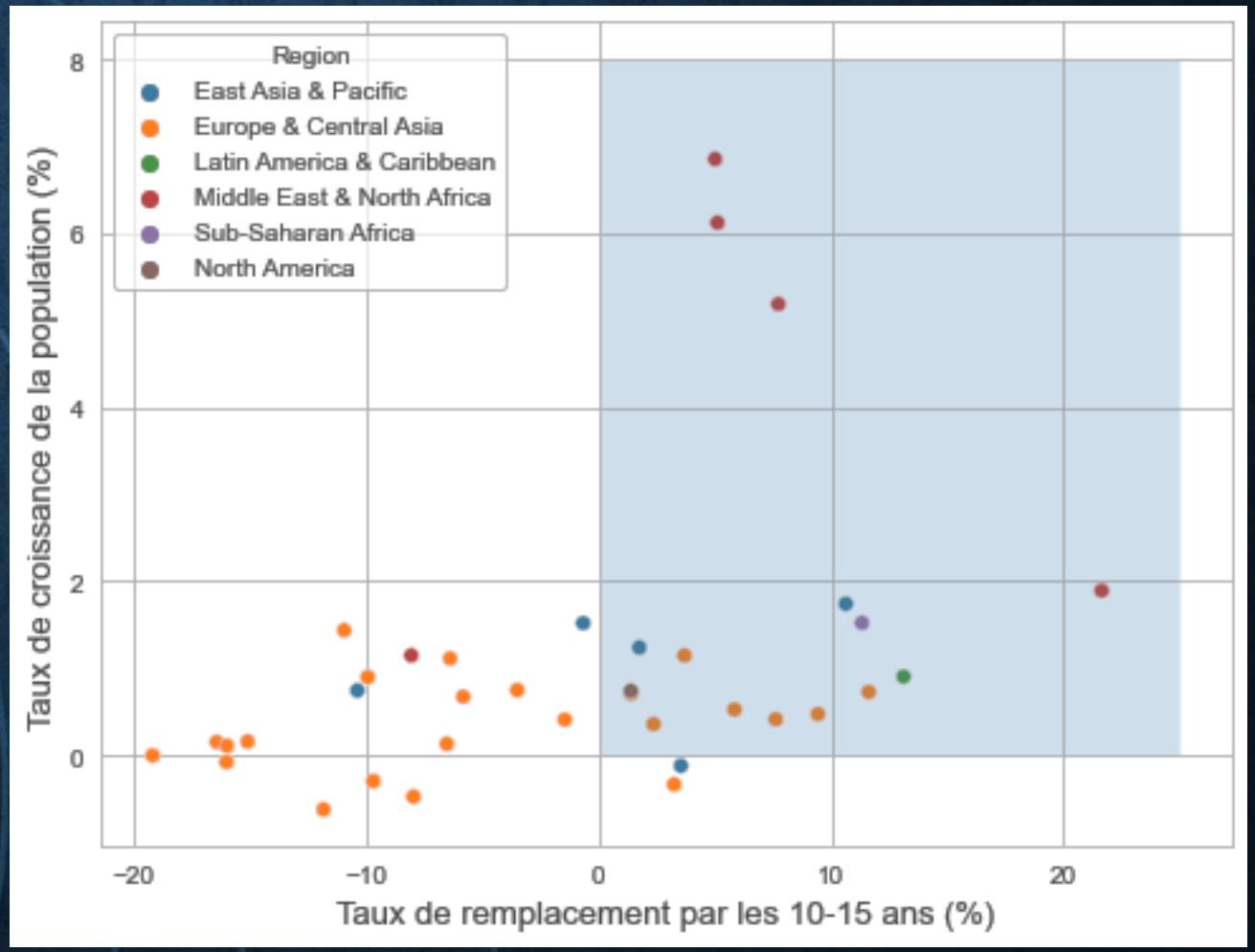
ETAPE 8 : ANALYSER LES DONNÉES

TRI N°3:

Taux de croissance de la population vs taux de remplacement par les 10-15 ans

RESULTATS:

« On passe d'une liste de 36 pays candidats à 16 pays remplissant tout les critères (dont fait partie la France) »



ETAPE 8 : ANALYSER LES DONNÉES

Région	Nb pays
Europe	6
Moyen-Orient	4
Asie, Pacifique	2
Amérique latine	1
Amérique du nord	1
Afrique sub-saharienne	1

ETAPE 9 : RÉPONDRE À L'OBJECTIF

	Nom pays	Classement prospection sur 6 critères
0	Norway	1
1	Netherlands	2
2	Denmark	3
3	Belgium	4
4	Israel	5
5	United States	6
6	New Zealand	7
7	Ireland	8
8	Qatar	9
9	Kuwait	10
10	Malaysia	11
11	Italy	12
12	Brazil	13
13	Oman	14
14	South Africa	15

LISTE DES CRITERES:

- **Taux accès internet (%)**
- **Taux d'étudiants (%)**
- **Population 15-24 ans (log)**
- **PIB ppp (log)**
- **Croissance population (%)**
- **Replacement 10-15 ans (%)**

MERCI DE VOTRE ATTENTION