



academy

Soutenance projet 02 - Data Science

Analyse des données de systèmes éducatifs



Fadia ACHIR

06 09 2020

Problématique

- **Academy** est une **start-up de la EdTech**
- **Elearnings** : Contenus de formation de **niveau lycée et université**
- Objectif d'**expansion à l'international** 



- Quels sont les pays avec un fort potentiel de clients pour nos services ?
- Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise doit-elle opérer en priorité ?



Sommaire

1. Environnement & Importation des librairies
2. Description de jeux de données
3. Analyse pré-exploratoire des données
4. Sélection des indicateurs
5. Déterminer des ordres de grandeurs des indicateurs statistiques
6. Quels sont les pays avec un fort potentiel de clients pour nos services
7. Conclusion
8. Boîte à outils

02 Jeux de données « Edstats »

Education Statistics (EdStats)

About Indicators Queries Country Themes Tools Blog

Find an Indicator

Search Indicators

Available Indicators >

Country at a Glance

Select a Country


EdStats Query

Data Download

WHAT'S NEW

The World Bank's Work in the Education Sector: New data on FY2016 Education Projects (Mar. 2017)

PISA 2015 science, math, and reading data are available in EdStats (Nov. 2016)

View More > 

Explore by Theme

Learning Outcomes

Attainment

Equity

Expenditures

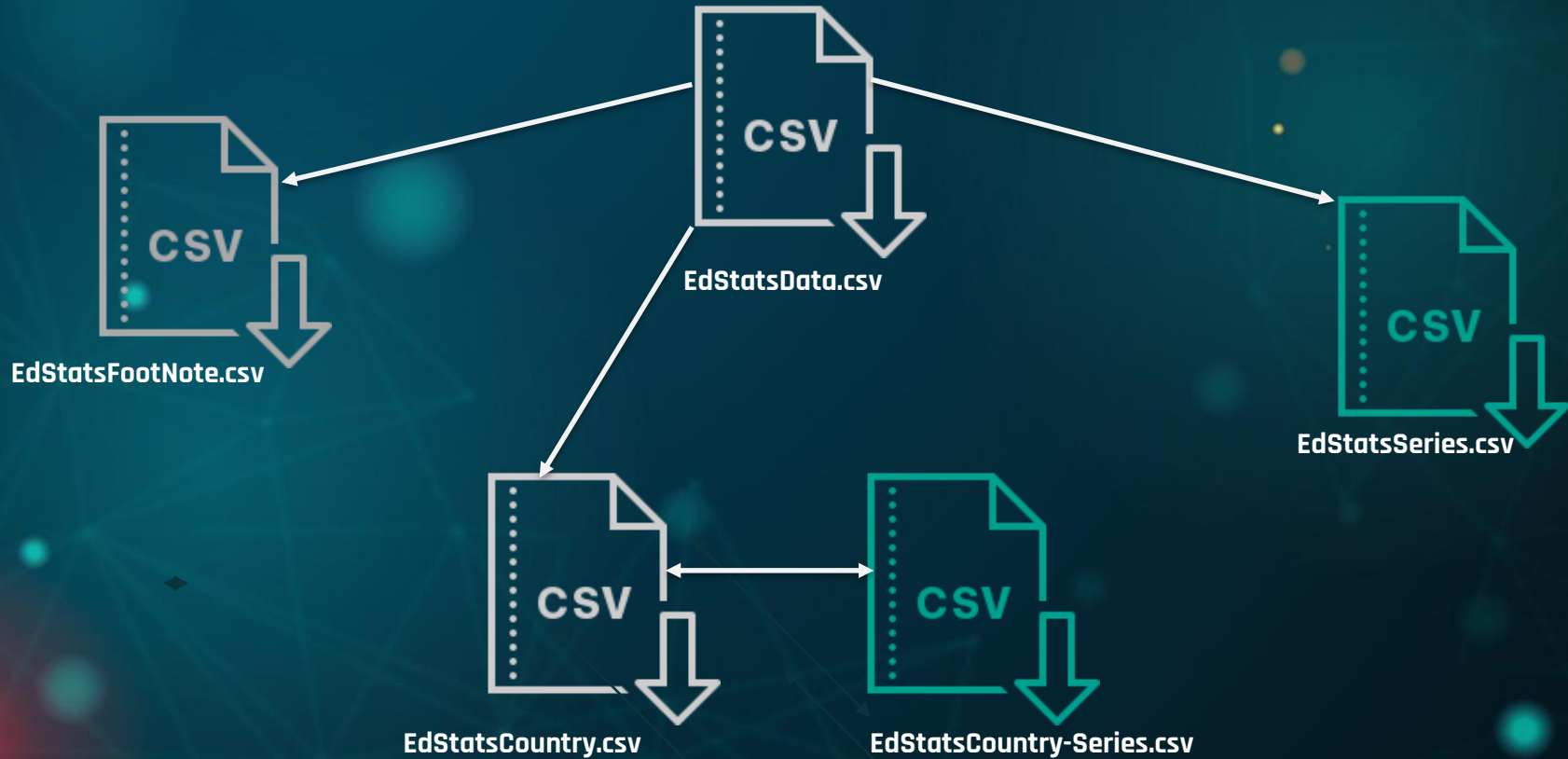
Policy

EMIS

Le portail “EdStats All Indicator Query” de la Banque mondiale répertorie 4000 indicateurs internationaux décrivant l'accès à l'éducation, l'obtention de diplômes et des informations relatives aux professeurs, aux dépenses liées à l'éducation, contenues dans 5 Datasets,

→ Pour en savoir plus :
<http://datatopics.worldbank.org/education/>

→ Site de la Banque Mondiale de données :
<http://datatopics.worldbank.org/education/>



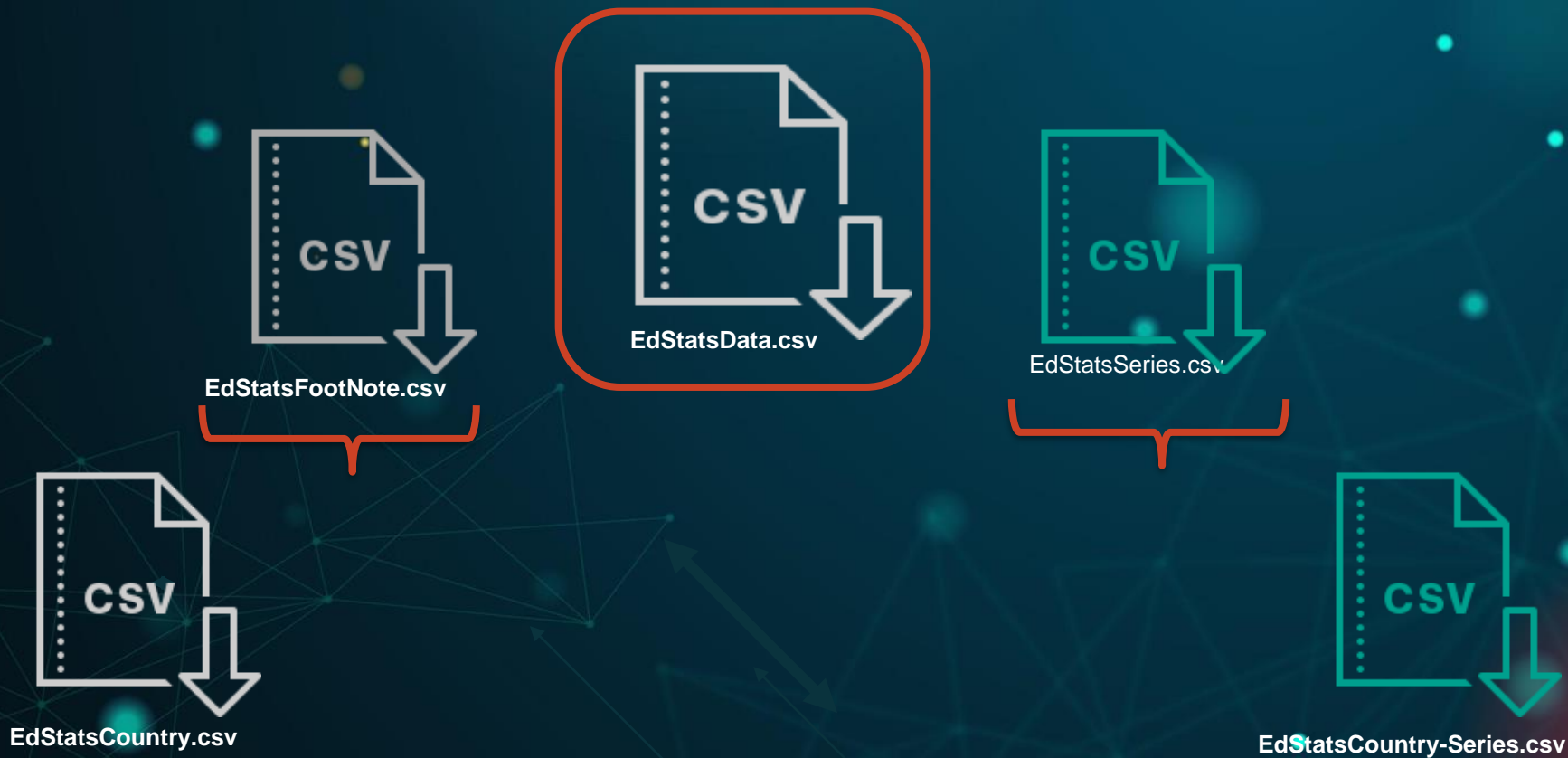
02|Objectif

- ✓ Valider la qualité de jeux de données;
- ✓ Décrire les informations contenues dans les jeux de données .
- ✓ Sélectionner les informations qui semblent pertinentes pour répondre à la problématique.
- ✓ Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique)

02 Description de données

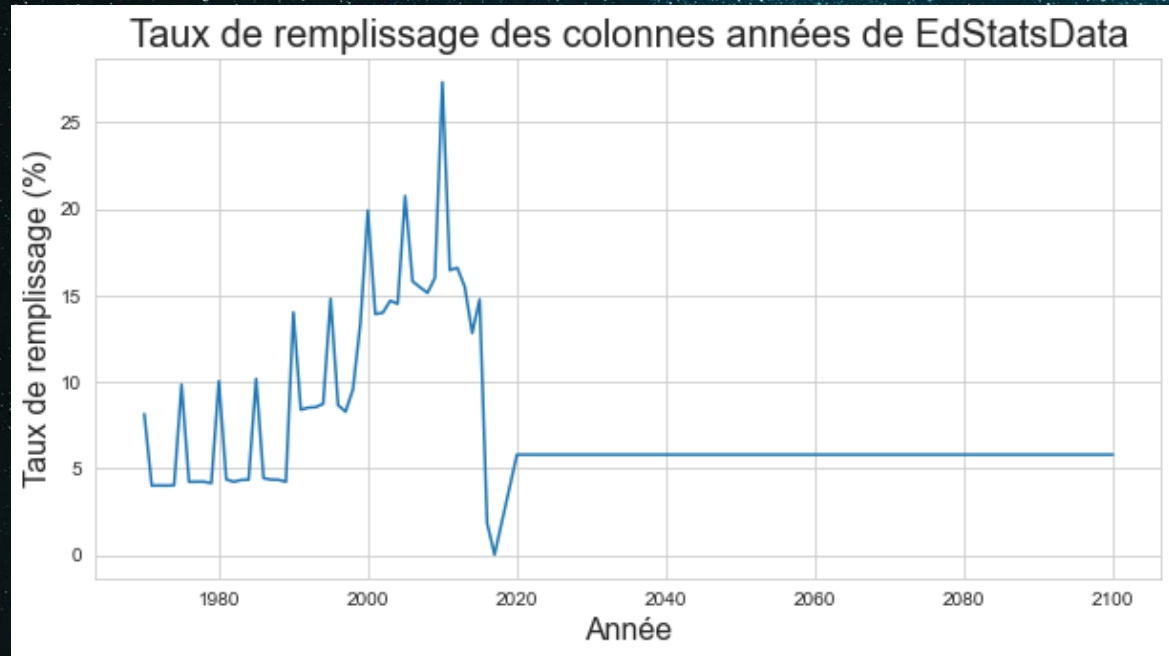
EdStatsCountry.csv	EdStatsCountry-Series.csv	EdStatsData.csv	EdStatsFootNote.csv	EdStatsSeries.csv
Des informations globales sur l'économie de chaque pays du monde (et de zones géographiques)	Le dataset contient les références des sources de certains indicateurs par pays dans la colonne (Description) présents dans le jeu de données EdStatsCountry.csv.	donne l'évolution de plusieurs indicateurs sur une période de 1970 à 2017 pour tous les pays du monde et certaines macros régions du monde, avec des prédictions pour les années 2020 - 2100	Le jeu de données contient les années de références de mises à jour des indicateurs par pays et la description des incertitudes, exceptions, remarques sur les mises à jour.	Le jeu de données permet de connaître le thème des indicateurs, les descriptions longues et les sources.
241 lignes et 32 colonnes	613 lignes et 4 colonnes	886 930 lignes, 70 colonnes	643638 lignes et 5 colonnes	3665 lignes et 21 colonnes
Pas de valeur manquante (sauf Unnamed : 31" qui est une colonne uniquement composée de NaN)	Pas de valeur manquante (sauf Unnamed : 3" qui est une colonne uniquement composée de NaN)	53455179 valeur manquantes près de 86%	Pas de valeur manquante (sauf Unnamed : 4 qui est une colonne uniquement composée de NaN)	Pas de valeur manquante (sauf Unnamed :20 qui est une colonne uniquement composée de NaN)
Aucun doublon	Aucun doublon	Aucun doublon	Aucun doublon	Aucun doublon

Suppression des datasets inutiles



Supprimer les colonnes Vides / NaN

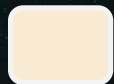
- 241 pays différents
- 3665 indicateurs différents
- Données de 1970 à 2017
- Prévisions de 2020 à 2100
- Nan : 86 %



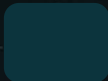
Supprimer les colonnes Vides / NaN

```
# supprime la colonne 'Unnamed: 69'
```

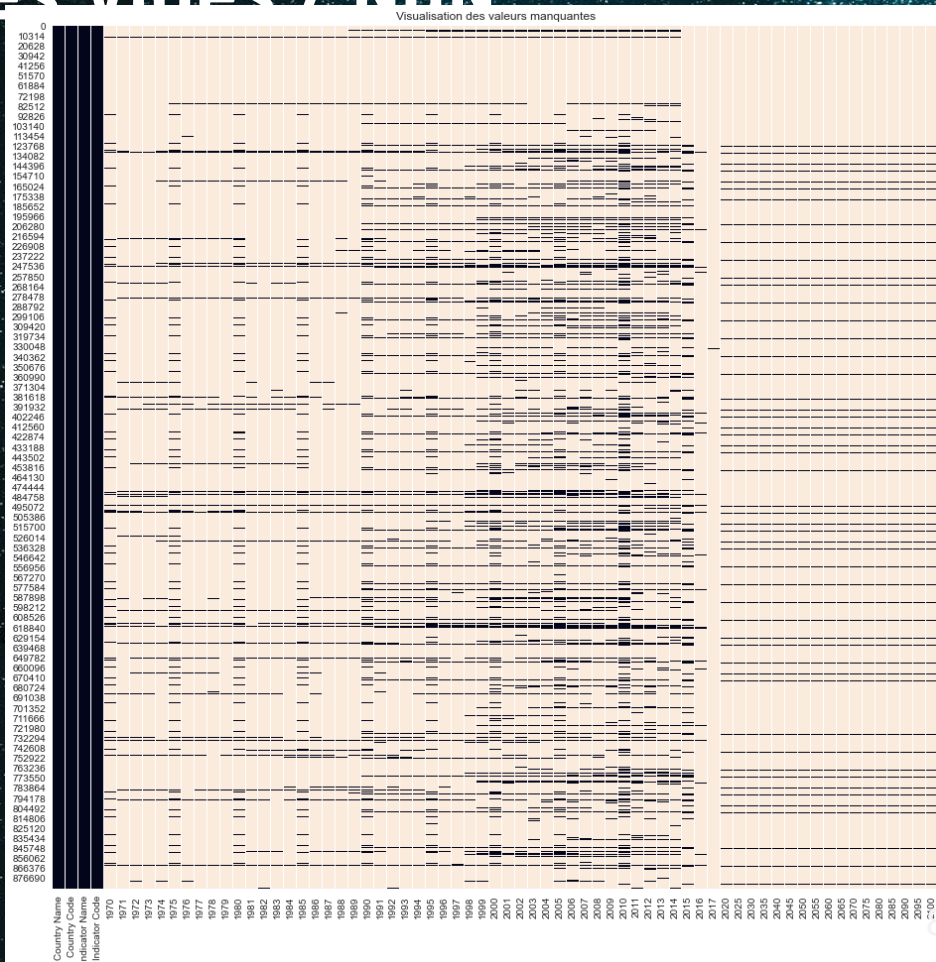
```
Df_data = Df_data.drop(['Unnamed: 69'], axis = 1)  
Df_data
```



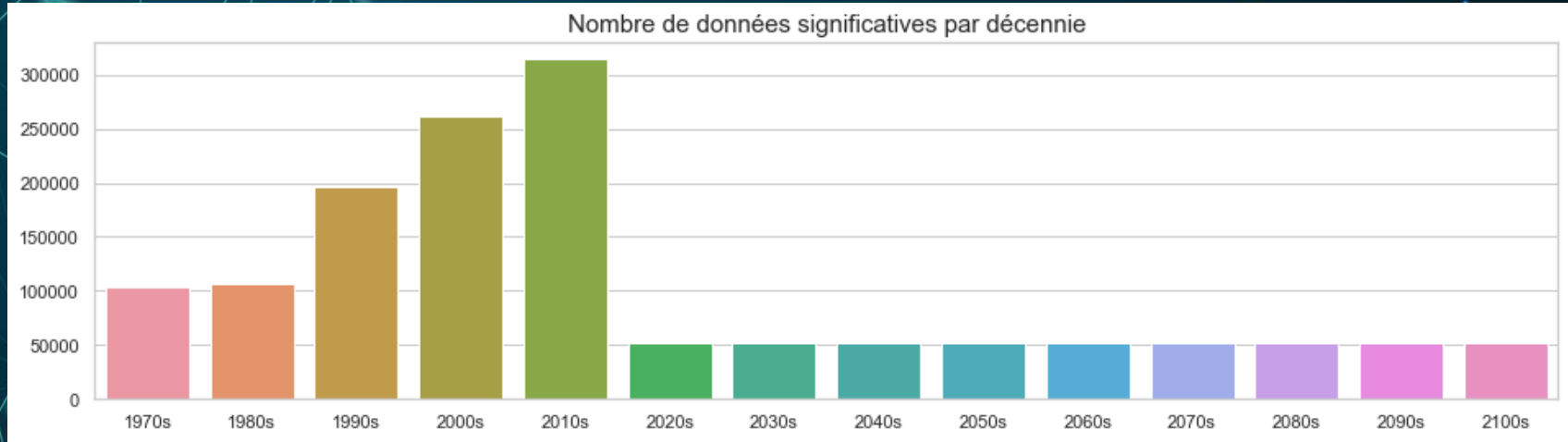
valeurs manquantes



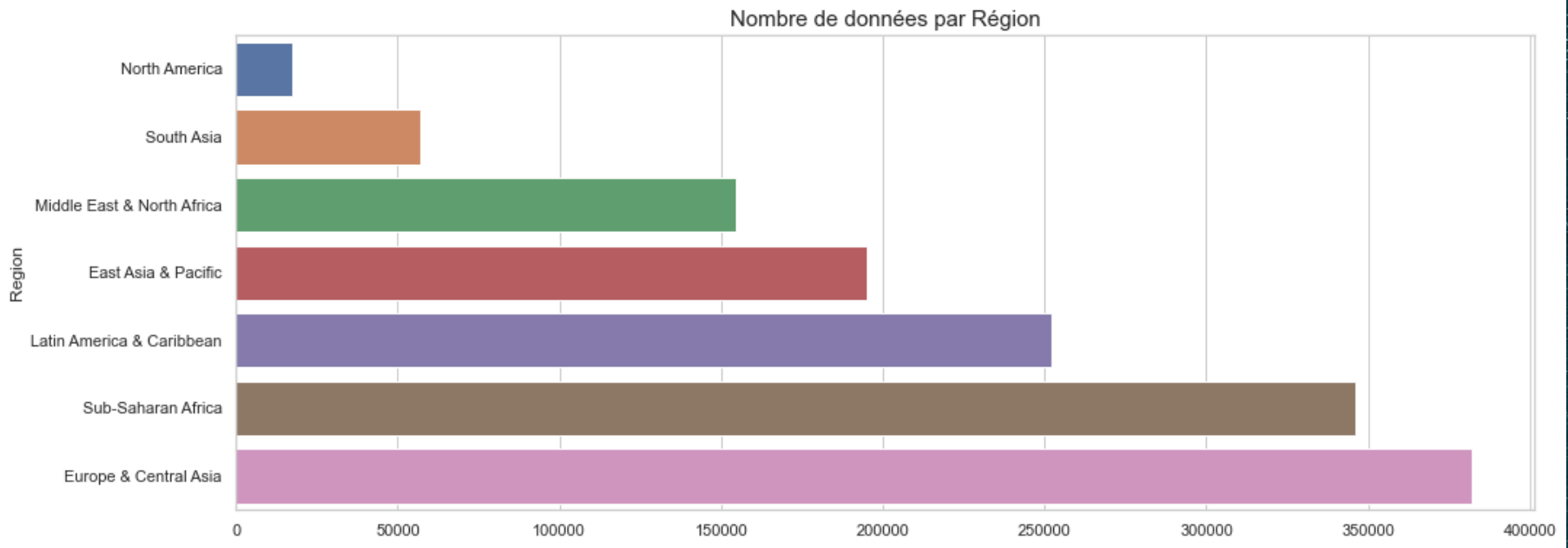
données



- un regroupement par décennies permet de voir que procéder à une analyse sur la décennie 2010's devrait permettre de procéder à une première analyse.



- On va procéder à notre analyse en utilisant les données de data enrichies de l'information "Region" pour faire des analyses d'indicateurs par groupes de pays.



Choix d'indicateurs



Forte connectivité à
internet



Indicateur des revenu
moyen PIB



La population estudiantine
(Lycéens et universitaires)

03 Indicateurs sélectionnés

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

Etude de l'indicateur de connectivité à internet : « IT.CMP.PCMP.P2 »

	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

Nous allons retirer l'indicateur " IT.CMP.PCMP.P2 " de la liste car il ne possède pas de données pour la décennie 2010's

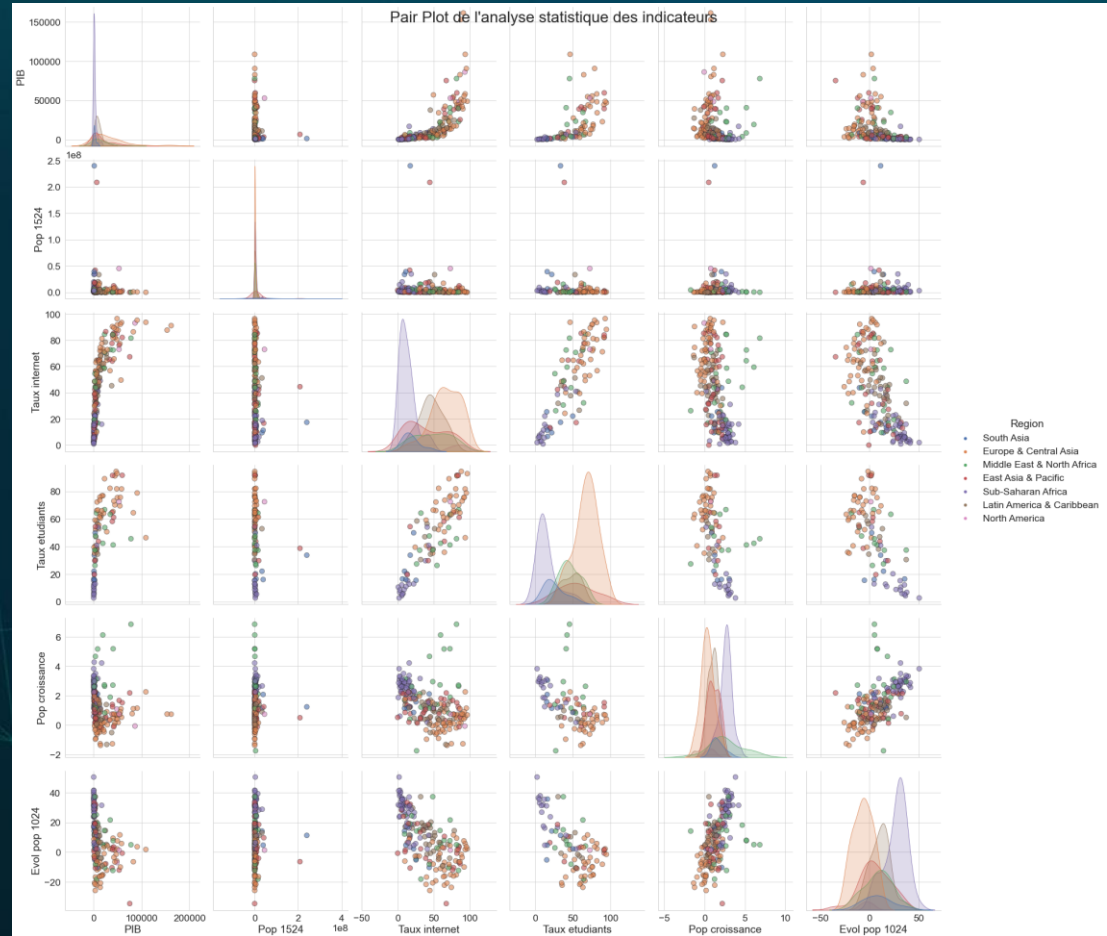
```
Selected_indicators.remove('IT.CMP.PCMP.P2')
```

Analyse Statistique :

	decennie courante	Taux internet	PIB	Pop 1015	Pop 1524	Pop totale	Pop croissance	Evol pop 1024	Pop etudiants	Taux etudiants
Taux valeur %	94.859813	94.859813	95.327103	84.579439	84.579439	100.0	100.0	84.579439	45.794393	42.056075
Moyenne	42.640728	42.640728	16456.621984	4011687.482505	6679532.875691	33458502.768243	1.362131	9.155498	3465679.169648	48.907839
Std	28.146515	28.146515	24747.710889	14156573.399946	24342092.500012	131917714.146908	1.319072	16.877185	11993738.727627	25.148724
Minimum	0.0	0.0	276.981192	8817.5	13861.833333	10815.571429	-1.744924	-34.494107	909.5	2.742091
Pays min	Korea, Dem. People's Rep.	Korea, Dem. People's Rep.	Burundi	Aruba	Aruba	Tuvalu	Syrian Arab Republic	Macao SAR, China	Turks and Caicos Islands	Niger
Mediane	42.912731	42.912731	6172.729261	814890.0	1306930.5	6232627.357143	1.202917	7.61065	485293.85	52.589627
Maximum	96.509472	96.509472	161621.007752	148501284.333333	240162812.666667	1357723571.428571	6.851502	50.472055	81031926.0	94.476159
Pays max	Iceland	Iceland	Liechtenstein	India	India	China	Qatar	Niger	India	Finland
Pays min	Korea, Dem. People's Rep.	Korea, Dem. People's Rep.	Burundi	Aruba	Aruba	Tuvalu	Syrian Arab Republic	Macao SAR, China	Turks and Caicos Islands	Niger

Analyse Statistique des indicateurs

1. Sur la diagonale nous obtenons les distribution de chacun de nos indicateurs
2. Dans la matrice inferieure et supérieure, nous avons l'analyse bivariée de deux indicateurs.



Analyse Statistique des indicateurs

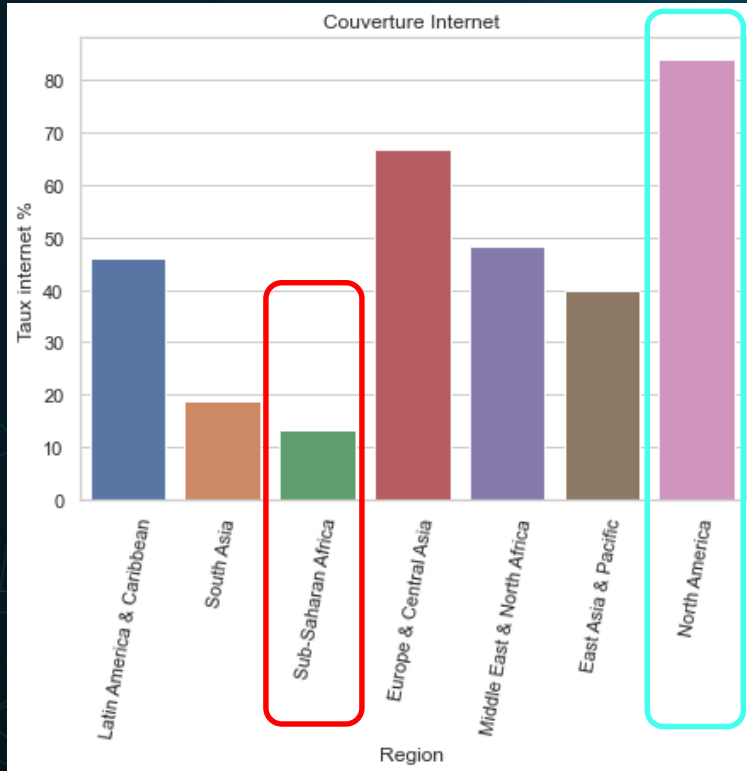
FONCTION HEATMAP :

→ Affiche en couleur l'indicateur de corrélation de Pearson r $[-1, 1]$ sur une matrice inférieure.

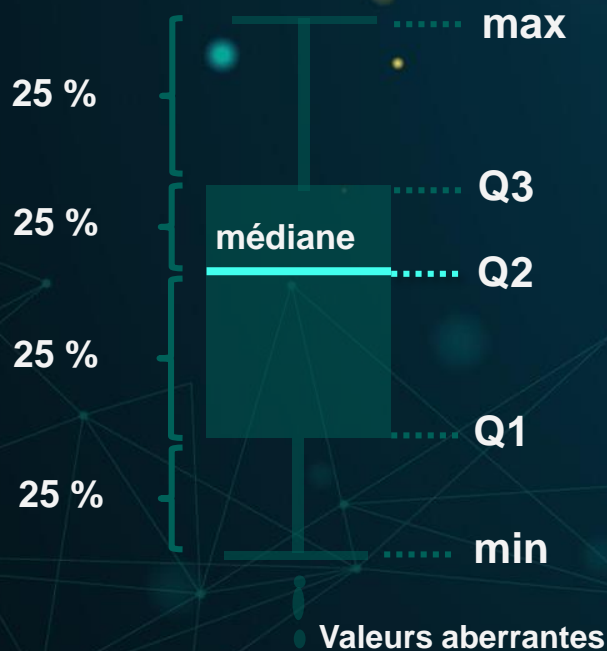
→ La **matrice de corrélation** montre une **dépendance** entre plusieurs indicateurs.



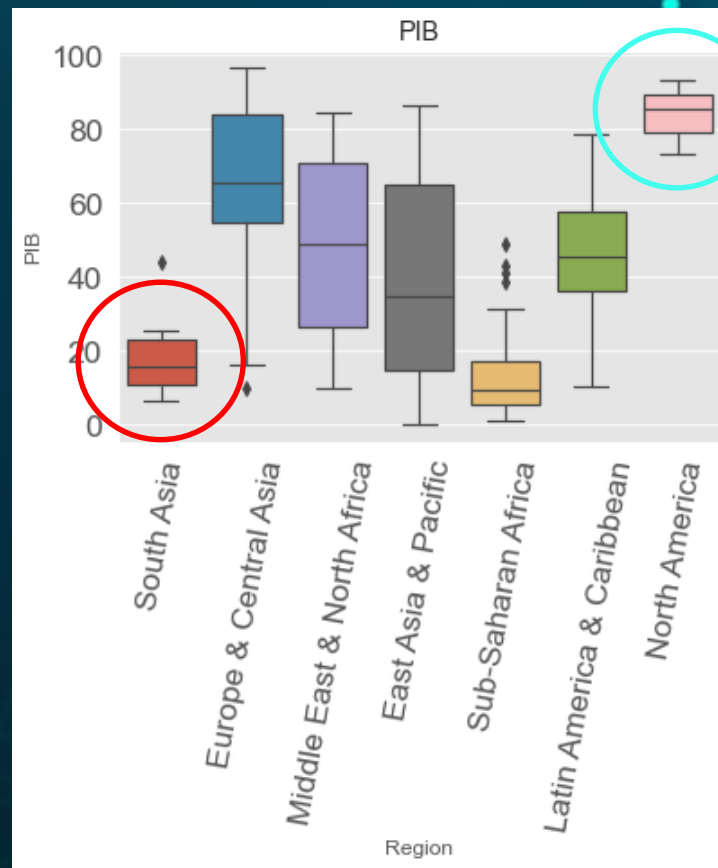
Etude de l'indicateur de connectivité à internet : « IT.Net.User.P2 »



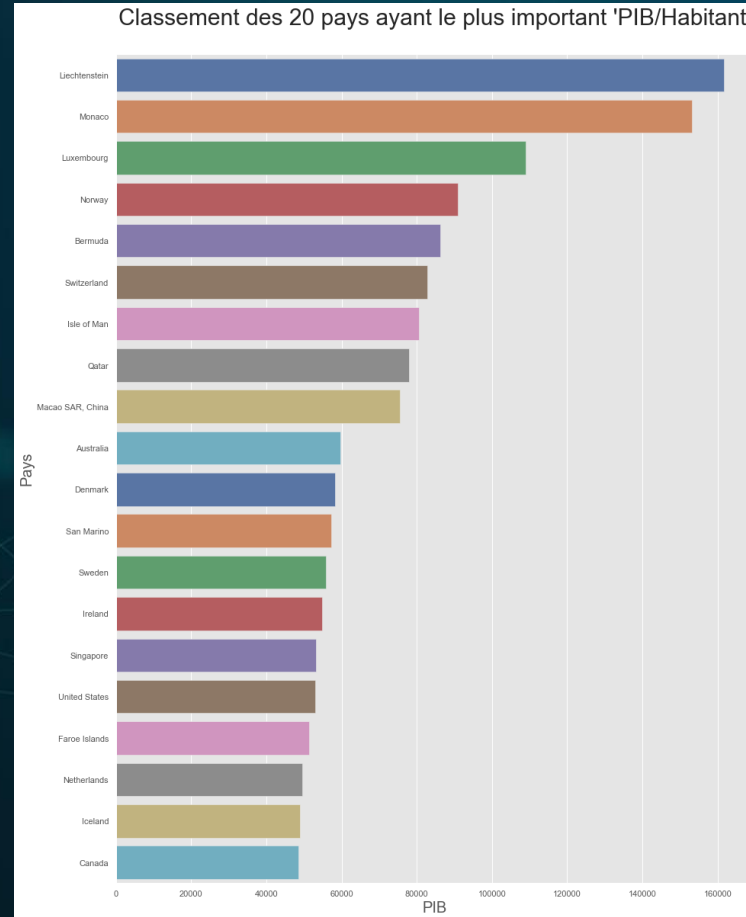
Etude de l'indicateur de richesse : PIB/habitant



Boîte à moustaches



Etude de l'indicateur de richesse : PIB/habitant



Choix des pays où Academy doit opérer en priorité

1. les pays les plus aisés

IT.NET.USER.P2

> 60%

NY.GDP.PCAP.CD

> 40000 \$/ habitant

SP.POP.1524.TO.UN

>1000

les pays avec une forte population

IT.NET.USER.P2

> 15 %

NY.GDP.PCAP.CD

> = 1000 \$/ pers

SP.POP.1524.TO.UN

> = 100000000


```
df_country_2010[((df_country_2010['Taux internet'] >= 15) & (df_country_2010['Pop 1524'] >= 1000) & (df_country_2010['PIB'] >= 1000)).sort_values(by='Pop 1524', ascending = False)[['Nom pays']].head(20)
```

#on va multiplier ce nombre avec le taux de pénétration d'internet pour avoir une estimation du nombre de clients potentiels:


```
df_country_2010['Final_score'] = df_country_2010['student_score'] * df_country_2010['Taux internet']/100
```


Projection :

Find an Indicator

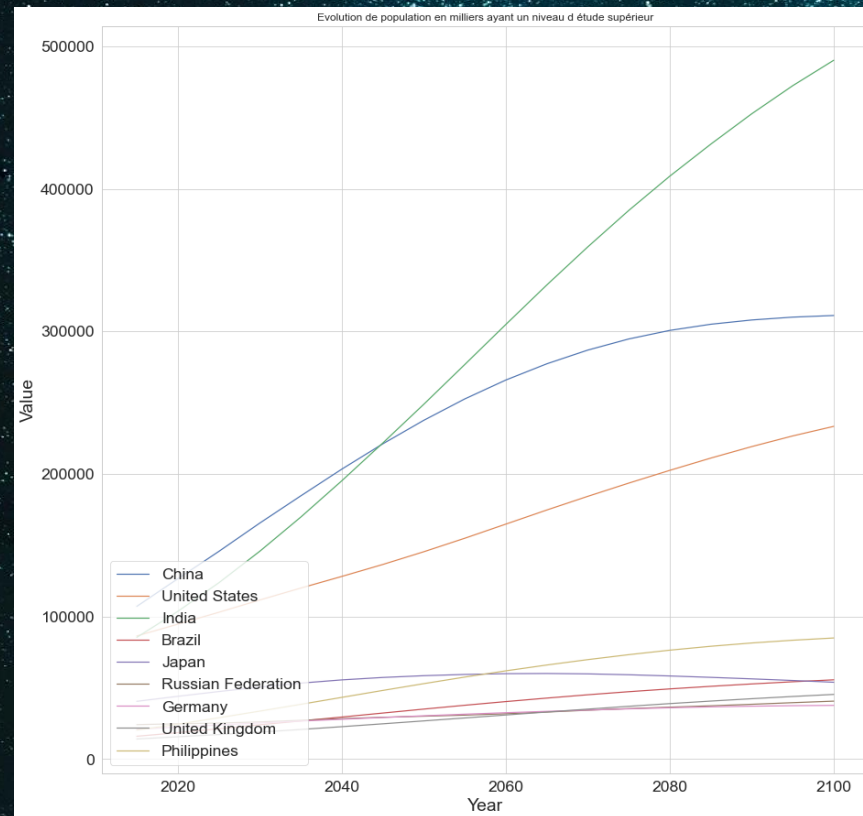
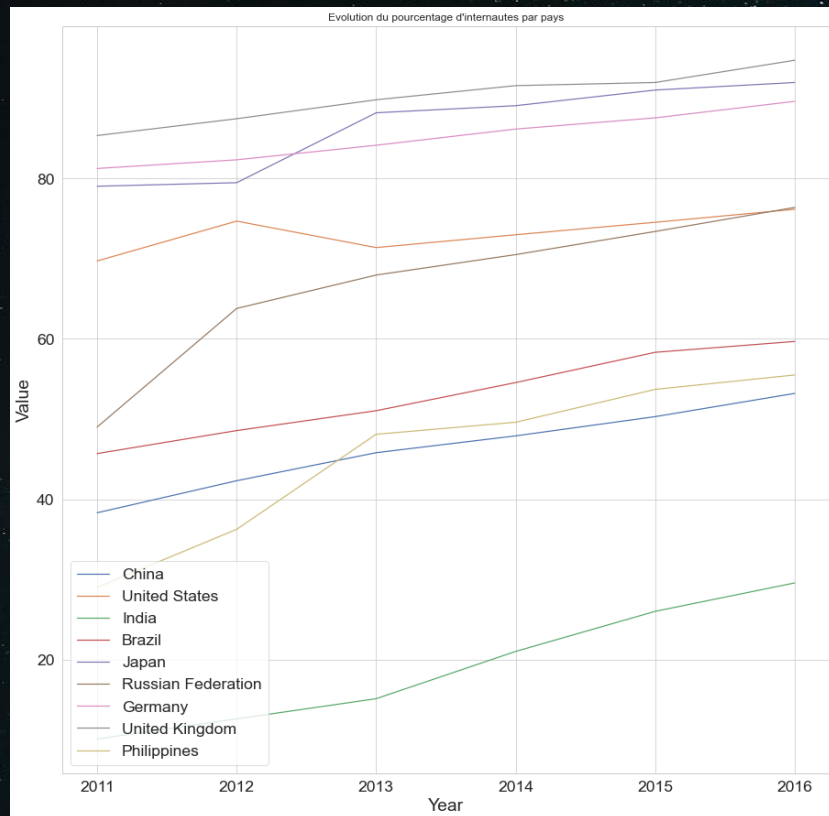
PRJ.POP.ALL.4.MF 

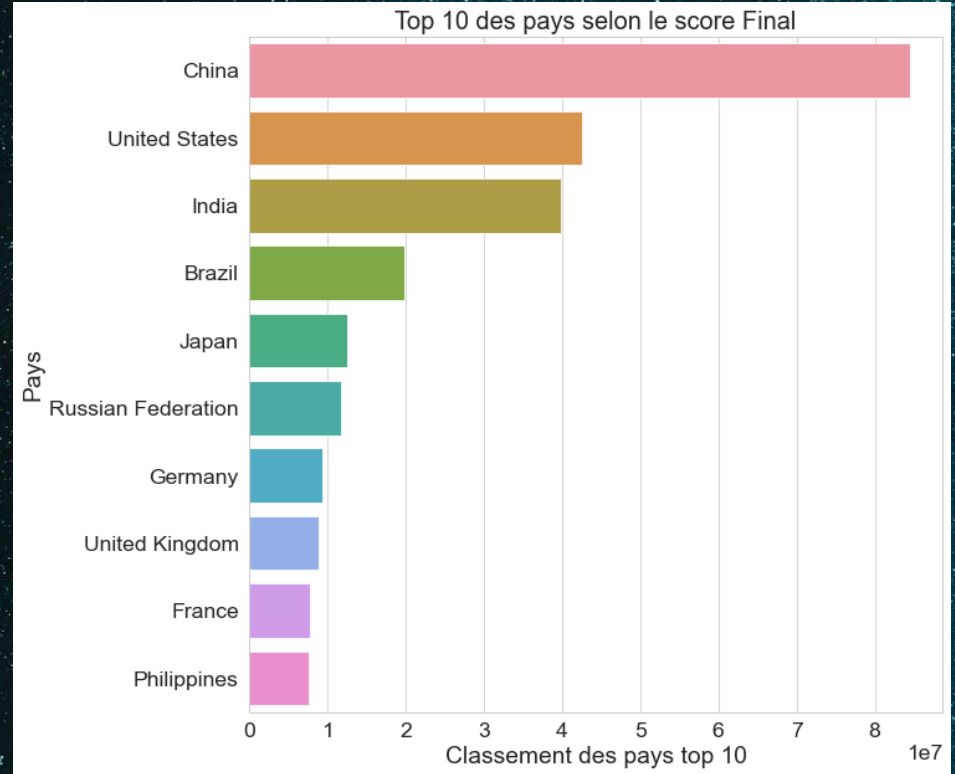
Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total

Select a Country 

EdStats Query 

Projection :





Bilan sur le jeux de données en termes de réponse aux attentes de Academy

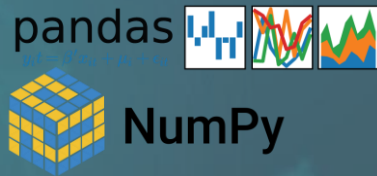
Qualités :	Limites :
Contient tous les pays du monde	Certains indicateurs inutilisables
Détaille plein de données relatives à l'éducation	Beaucoup de données manquantes pour comparer
Mentionne la source de la données	Manque certains indicateurs business
Pertinent pour répondre à la problématique	Manque d'informations sur la société Academy

Boîte à outils

Environnement



Librairies de base



Visualisation



Data exploration



Merci!

