OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Soutenance du Projet 2

Analysez des données de systèmes éducatifs

Par
Elisée TCHANA

Mentor
Cyril MONTI

03 Févr 2021

OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Plan de Soutenance

Problématique et Présentation du jeu de données

+

Analyse pré-exploratoire du jeu de données et Conclusions sur la pertinence de l'usage du jeu de données pour répondre aux questions stratégiques que se pose l'entreprise

→

Questions-réponses

Objectifs du Projet

Objectifs

- Cibler les pays à fort potentiel pour une implantation Academy EdTech
- Observer les potentielles évolutions de ces pays
- Déterminer dans quels pays opérer en priorité

Hypothèse du Projet

1. Valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)
2. Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)
3. Sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise ?)
4. 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)

Problématique

Expansion à l'international pour une Startup de la Edtech

OPENCLASSROOMS

Data Analysis Workflow

CentraleSupélec | université PARIS-SACLAY

BANQUE MONDIALE

EdStatsCountry

EdStatsCountry-Series

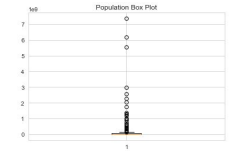
EdStatsData

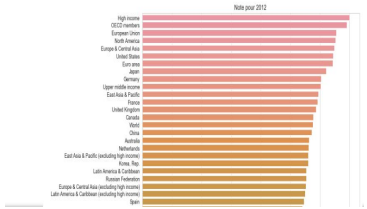
EdStatsFootNote

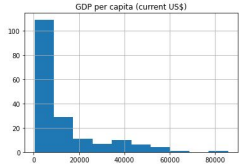
EdStatsSeries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math
from scipy import interpolate
import plotly.express as px
data= pd.read_csv("C:/Users/Elisée/Desktop/Projet (2)/Projet+Python_Dataset_Edstats_csv/EdStatsData.csv")
raw_data = data.copy()
country = pd.read_csv("C:/Users/Elisée/Desktop/Projet (2)/Projet+Python_Dataset_Edstats_csv/EdStatsCountry.csv")
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary,	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN







Data Analysis

Data Sharing

OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Informations contenues dans le jeu de données (Pays, Indicateurs, Années)

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	20
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	N
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	N
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	N
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	N
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN	NaN	NaN	NaN	NaN	N

Caractéristiques de la base de données:
886930 lignes
70 colonnes

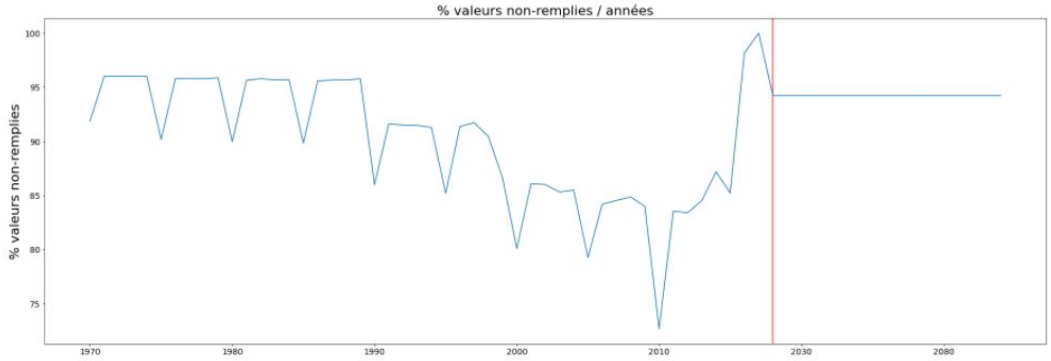
242 pays différents
3665 indicateurs différents

```
Adjusted net enrolment rate, lower secondary, both sexes (%)
Adjusted net enrolment rate, lower secondary, female (%)
Adjusted net enrolment rate, lower secondary, gender parity index (GPI)
Adjusted net enrolment rate, lower secondary, male (%)
Adjusted net enrolment rate, primary, both sexes (%)
Adjusted net enrolment rate, primary, female (%)
Adjusted net enrolment rate, primary, gender parity index (GPI)
Adjusted net enrolment rate, primary, male (%)
Adjusted net enrolment rate, upper secondary, both sexes (%)
Adjusted net enrolment rate, upper secondary, female (%)
Adjusted net enrolment rate, upper secondary, gender parity index (GPI)
Adjusted net enrolment rate, upper secondary, male (%)
Adjusted net intake rate to Grade 1 of primary education, both sexes (%)
Adjusted net intake rate to Grade 1 of primary education, female (%)
Adjusted net intake rate to Grade 1 of primary education, gender parity index (GPI)
Adjusted net intake rate to Grade 1 of primary education, male (%)
Adult illiterate population, 15+ years, % female
Adult illiterate population, 15+ years, both sexes (number)
Adult illiterate population, 15+ years, female (number)
```

4000 indicateurs internationaux décrivant l'accès à l'éducation

Nettoyage des données

L'ensemble de données contient 95% de “nan” par années mais une diminution vers 2000




```
data=data.reindex(columns=['Country Name','Country Code','Indicator Name','Indicator Code',
                           '2010','2011','2012','2013','2014','2015'])
data.head()
```

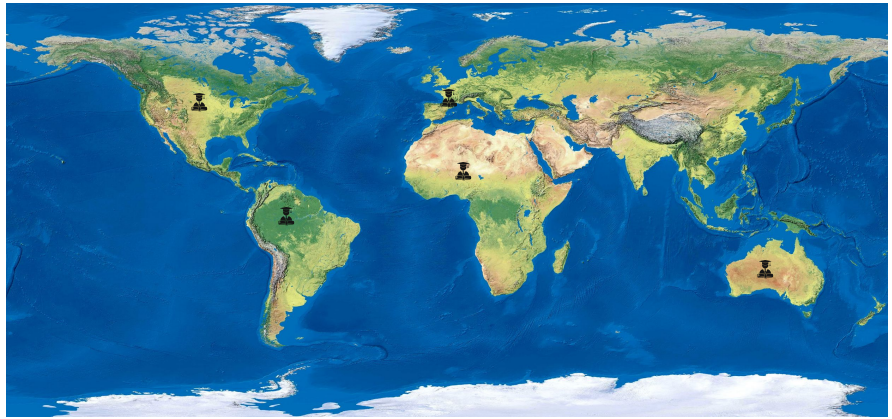
	Country Name	Country Code	Indicator Name	Indicator Code	2010	2011	2012	2013	2014	2015
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	85.211998	85.24514	86.101669	85.51194	85.320152	NaN

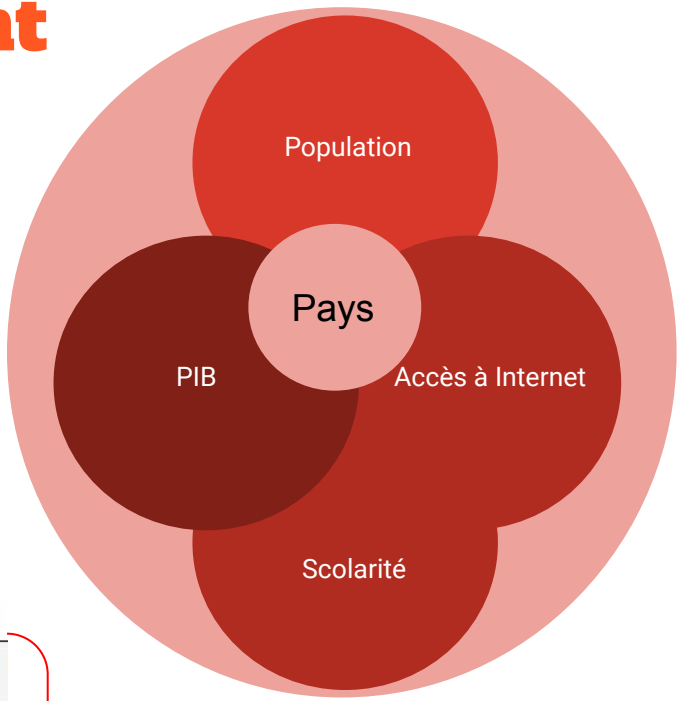
Notre analyse s'est porté sur une décennie récente sur une durée de Cinq ans. Cette décennie nous aide à définir une meilleure stratégie pour la Edtech

OPENCLASSROOMS

CentraleSupélec | université PARIS-SACLAY

Indice de développement





4000 indicateurs internationaux

60%
Moyenne

Indicator Name
GDP per capita (current US$)
Gross enrolment ratio, secondary, both sexes (%)
Internet users (per 100 people)
Population, total
GDP per capita (current US$)

242 pays différents → 216 pays restants

Avant **Après**

Data Cleaning & Data Analysis

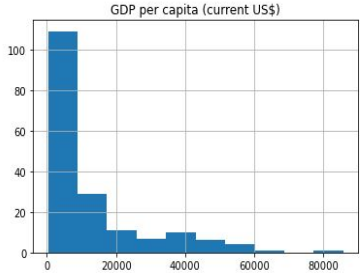
Analyse des Nans → Méthode de remplissage manuel par les valeurs de la cellule voisine

```
for ind in range(0,len(relevant_indicators)):
    print(relevant_indicators[ind],": \n",
          data_manual[data_manual['Indicator Name'] == relevant_indicators[ind]].describe()," \n---------- \n")
```

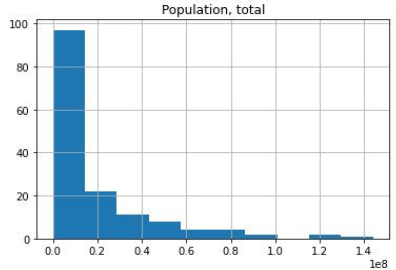
```
Population, total :
              2010          2011          2012          2013          2014  \
count  1.840000e+02  1.840000e+02  1.840000e+02  1.840000e+02  1.840000e+02
mean   2.515622e+08  2.546393e+08  2.577727e+08  2.609632e+08  2.640943e+08
std    8.562833e+08  8.665048e+08  8.769007e+08  8.874179e+08  8.979575e+08
min    1.053100e+04  1.062800e+04  1.072500e+04  1.081900e+04  1.090800e+04
25%    3.030616e+06  3.150317e+06  3.293610e+06  3.295426e+06  3.297751e+06
50%    1.071705e+07  1.089832e+07  1.096584e+07  1.103985e+07  1.117648e+07
75%    5.305393e+07  5.364187e+07  5.426482e+07  5.504245e+07  5.580734e+07
max    6.930657e+09  7.012844e+09  7.097401e+09  7.182860e+09  7.268986e+09

              2015          Mean
count  1.840000e+02  1.840000e+02
mean   2.672338e+08  2.216271e+08
std    9.085239e+08  7.557216e+08
min    1.100100e+04  9.230286e+03
25%    3.317883e+06  2.826996e+06
50%    1.127393e+07  9.404996e+06
75%    5.644163e+07  4.689315e+07
max    7.355220e+09  6.121138e+09
----------
```

Tableau nominative pour chaque indicateur

Représentation de la population avec un boxplot pour montrer les valeurs aberrantes



Echelle Logarithmique



Echelle Logarithmique

Représentant la répartition du PIB et de la Population des pays sans grandes valeurs aberrantes

OPENCLASSROOMS

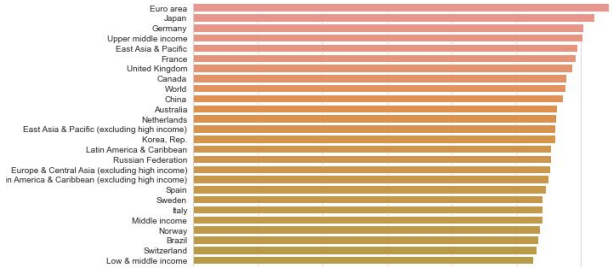
CentraleSupélec | université PARIS-SACLAY

Data Analysis

Hiérarchisation des pays

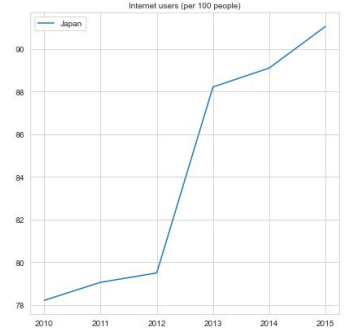
```
first_test_grade = data_manual.reindex(columns=['Country Name','2010 Grade','2011 Grade','2012 Grade','2013 Grade','2014 Grade',
first_test_grade = first_test_grade.groupby(['Country Name']).sum()
first_test_grade.head(50)
```

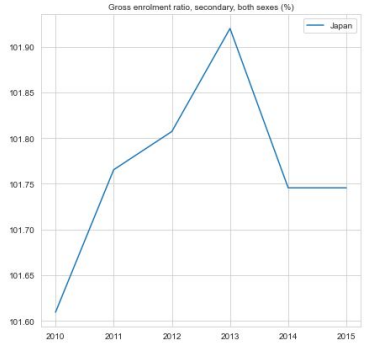
Country Name	2010 Grade	2011 Grade	2012 Grade	2013 Grade	2014 Grade	2015 Grade
Afghanistan	5.692514	5.818627	5.952393	6.004232	6.039163	6.118895
Albania	7.674336	8.000684	8.262095	8.519637	8.690122	8.651935
Algeria	9.350770	9.748751	9.922564	10.100008	10.399501	10.523782
Angola	5.577156	5.805751	6.037715	6.208570	6.278020	6.184919
Antigua and Barbuda	8.105696	8.287706	8.570428	8.763152	8.991665	9.128149
Arab World	11.317227	11.695683	11.998809	12.092332	12.110746	12.110746
Argentina	11.640405	12.081448	12.300913	12.475992	12.626789	12.855434
Aruba	9.153053	9.518527	9.837922	10.033922	10.229122	10.424371
Australia	13.687727	14.003609	14.076060	14.272895	14.233225	14.184784



	Country Name	Country Code	Indicator Name	Indicator Code	2010	2011	2012	2013	2014	2015	Mean
448375	Japan	JPN	GDP per capita (current US$)	NY.GDP.PCAP.CD	4.450768e+04	4.816800e+04	4.860348e+04	4.045445e+04	3.809621e+04	3.447414e+04	3.632914e+04
448465	Japan	JPN	Gross enrolment ratio, secondary, both sexes (%)	SE.SEC.ENRR	1.016095e+02	1.017654e+02	1.018073e+02	1.019200e+02	1.017456e+02	1.017456e+02	8.497463e+01
448505	Japan	JPN	Internet users (per 100 people)	IT.NET.USER.P2	7.821000e+01	7.905411e+01	7.949640e+01	8.821943e+01	8.910683e+01	9.105803e+01	7.216354e+01
449663	Japan	JPN	Population, total	SP.POP.TOTL	1.280700e+08	1.278330e+08	1.276290e+08	1.274450e+08	1.272760e+08	1.271410e+08	1.093420e+08

	2010	2011	2012	2013	2014	2015
1	High income	High income	High income	High income	High income	High income
2	OECD members	OECD members	OECD members	OECD members	OECD members	OECD members
3	European Union	European Union	European Union	European Union	Europe & Central Asia	Europe & Central Asia
4	North America	Europe & Central Asia	North America	Europe & Central Asia	European Union	North America
5	Euro area	Euro area	Europe & Central Asia	North America	North America	European Union
6	United States	North America	United States	Euro area	Euro area	United States
7	Europe & Central Asia	United States	Euro area	United States	United States	Euro area
8	Japan	Japan	Japan	Japan	Japan	Japan
9	Germany	Germany	Germany	Upper middle income	Upper middle income	Upper middle income
10	United Kingdom	Upper middle income	Upper middle income	Germany	Germany	East Asia & Pacific
11	France	France	East Asia & Pacific	East Asia & Pacific	United Kingdom	United Kingdom
12	Upper middle income	United Kingdom	France	United Kingdom	East Asia & Pacific	Germany
13	East Asia & Pacific	East Asia & Pacific	United Kingdom	France	France	China
14	Canada	Canada	Canada	China	China	World
15	World	World	World	World	World	France
16	Netherlands	Netherlands	China	Canada	Canada	East Asia & Pacific (excluding high income)
17	Korea, Rep.	Australia	Australia	East Asia & Pacific (excluding high income)	East Asia & Pacific (excluding high income)	Canada
18	Australia	Korea, Rep.	Netherlands	Australia	Korea, Rep.	Korea, Rep.
19	Spain	China	East Asia & Pacific (excluding high income)	Europe & Central Asia (excluding high income)	Europe & Central Asia (excluding high income)	Latin America & Caribbean
20	Italy	East Asia & Pacific (excluding high income)	Korea, Rep.	Korea, Rep.	Latin America & Caribbean	Latin America & Caribbean (excluding high income)





Hypothèse vérifiée

1. Valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?) ✅
2. Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?) ✅
3. Sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise ?) ✅
4. 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) ✅

Conclusion

	2010	2011	2012	2013	2014	2015
1	High income	High income	High income	High income	High income	High income
2	OECD members	OECD members	OECD members	OECD members	OECD members	OECD members
3	European Union	European Union	European Union	European Union	Europe & Central Asia	Europe & Central Asia
4	North America	Europe & Central Asia	North America	Europe & Central Asia	European Union	North America
5	Euro area	Euro area	Europe & Central Asia	North America	North America	European Union
6	United States	North America	United States	Euro area	Euro area	United States
7	Europe & Central Asia	United States	Euro area	United States	United States	Euro area
8	Japan	Japan	Japan	Japan	Japan	Japan
9	Germany	Germany	Germany	Upper middle income	Upper middle income	Upper middle income
10	United Kingdom	Upper middle income	Upper middle income	Germany	Germany	East Asia & Pacific
11	France	France	East Asia & Pacific	East Asia & Pacific	United Kingdom	United Kingdom
12	Upper middle income	United Kingdom	France	United Kingdom	East Asia & Pacific	Germany
13	East Asia & Pacific	East Asia & Pacific	United Kingdom	France	France	China
14	Canada	Canada	Canada	China	China	World
15	World	World	World	World	World	France
16	Netherlands	Netherlands	China	Canada	Canada	East Asia & Pacific (excluding high income)
17	Korea, Rep.	Australia	Australia	East Asia & Pacific (excluding high income)	East Asia & Pacific (excluding high income)	Canada
18	Australia	Korea, Rep.	Netherlands	Australia	Korea, Rep.	Korea, Rep.
19	Spain	China	East Asia & Pacific (excluding high income)	Europe & Central Asia (excluding high income)	Europe & Central Asia (excluding high income)	Latin America & Caribbean
20	Italy	East Asia & Pacific (excluding high income)	Korea, Rep.	Korea, Rep.	Latin America & Caribbean	Latin America & Caribbean (excluding high income)

On peut observer un classement des pays et régions sur les 6 années

La Edtech peut penser s'orienter vers les pays ayant une langue commune

Suivre les pays présentant une forte évolution sur ces années.

Merci pour votre attention !!