Analysez des données de systèmes éducatifs

Mitra Dadgar _ Data Scientist

1. Présentation du contexte
2. Présentation des données
3. Qualité des données
4. Selection des indicateurs
5. Analyse univariée
6. Scoring
7. Classement final
8. Evolution du potentiel
9. Conclusion

Présentation du contexte

- ***Academy*** **(start-up de la EdTech) :** propose des contenus de la formation en ligne niveau lycée et université.

- ***l'objectif*** : Analyser la possibilité d'un projet d'expansion à l'international de l'entreprise

- Ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays:

➔ 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 ?

Présentation des données

- **La source du dataset**: Banque mondiale
- **Nombre de lignes** : 886930 lignes
- **Nombre de colonnes :** 101 colonnes
- **Nombre de pays** : 1302 pays
- **Nombre d'indicateurs:** 3665 indicateurs

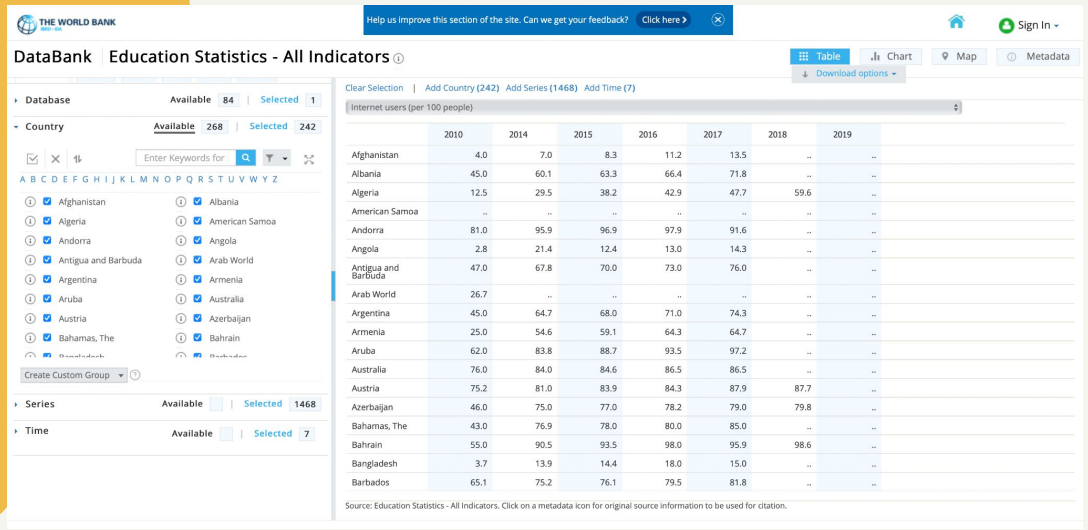
LA BANQUE MONDIALE
BIRD • IDA | GROUPE DE LA BANQUE MONDIALE

- **Avant nettoyage**

```
# Taille de notre dataframe
df.shape
```

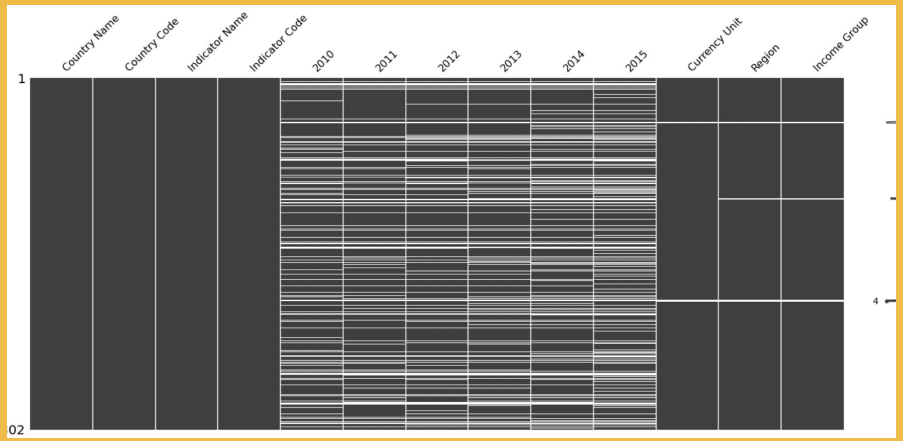
```
(886930, 101)
```

Comment sélectionner les indicateurs intéressants ?
sélectionner les indicateurs les mieux remplis



Qualité des données

- ➔ Comporte-t-il beaucoup de données manquantes?
- ➔ Existe-t-il des valeurs duplicates dans ce jeu de données ? Non
- ➔ Identifier les indicateurs en fonction de la qualité



- **Après nettoyage**

```
df2.shape
```

```
(875, 14)
```

Taux de valeurs manquantes par année

2017	99.983877
2016	98.144160
1971	95.993258
1973	95.992356
1972	95.984012
1974	95.971497
1979	95.849842
1976	95.773849
1982	95.770692
1989	95.767422
1977	95.763589
1978	95.763364
1983	95.663694
1988	95.653321
1984	95.647233
1987	95.643286
1981	95.627953
1986	95.560867
2020	94.200670
2025	94.200670
2030	94.200670
2035	94.200670
2040	94.200670
2045	94.200670
2050	94.200670
2055	94.200670
2060	94.200670
2065	94.200670
2070	94.200670
2075	94.200670
2080	94.200670
2085	94.200670
2090	94.200670
2095	94.200670
2100	94.200670
1970	91.849639
1997	91.718287
1991	91.607342
1992	91.482642
1993	91.454455
1996	91.340128
1994	91.266278
1998	90.426076
1975	90.156382
1980	89.951631
1985	89.819264
2014	87.170464

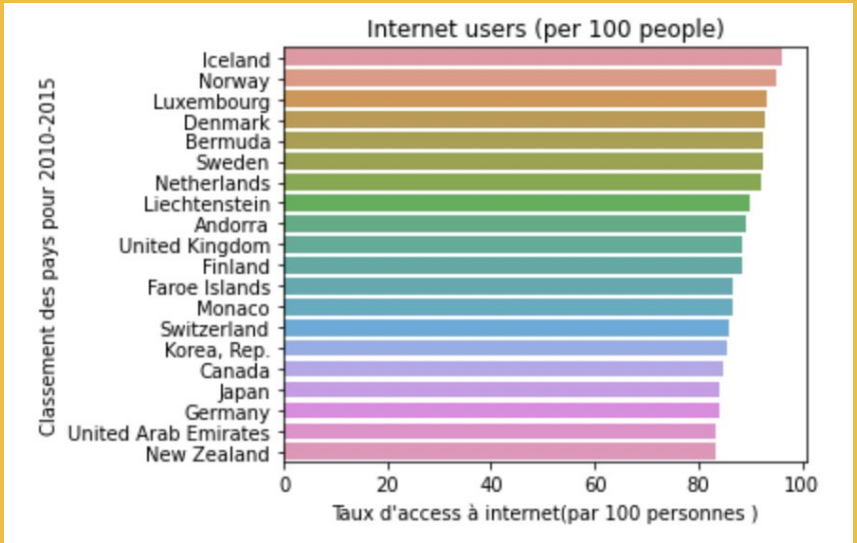
Selection des Indicateurs

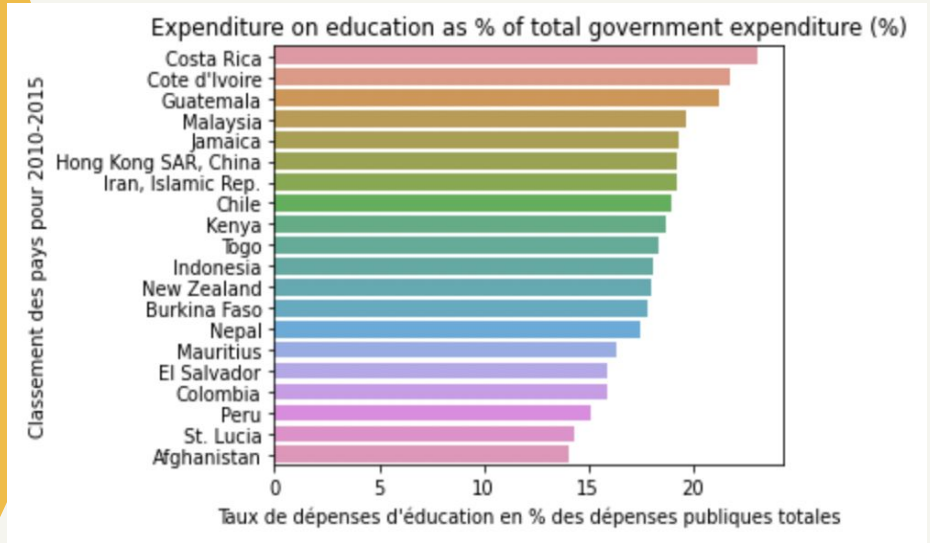
les indicateurs importants pour l'étude de marché :

- ➔ Quel est le nombre de clients potentiels dans le pays (ceux qui ont envie de se former) niveau lycée et université.
  - ❑ Enrolment in tertiary education, all programmes, both sexes (number)
  - ❑ Population, ages 15-24, total
  - ❑ Population of the official entrance age to secondary general education, both sexes (number)
  - ❑ Population growth (annual %)
- ➔ Les prospects peuvent-ils accéder à notre service (ordinateur/smartphone/internet)
  - ❑ Internet users (per 100 people)
- ➔ Les prospects peuvent-ils payer notre service
  - ❑ Expenditure on education as % of total government

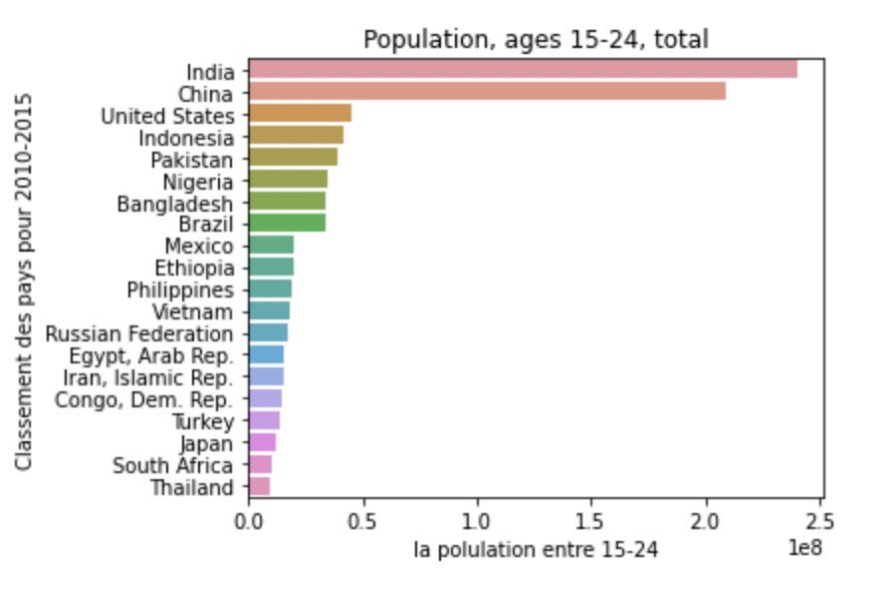
Accès à Internet et dépense dans l'éducation

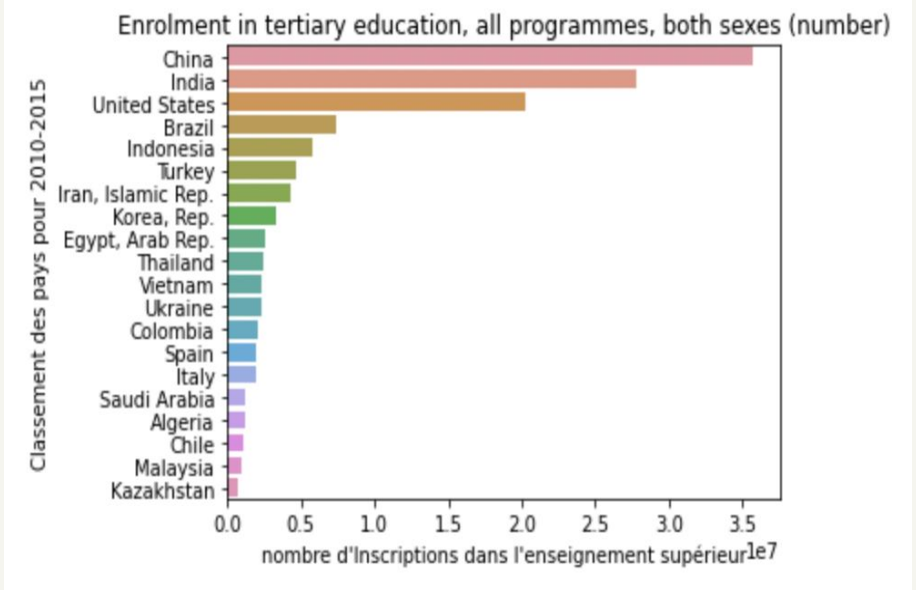
- Classement des pays par indicateurs pour 2010-2015



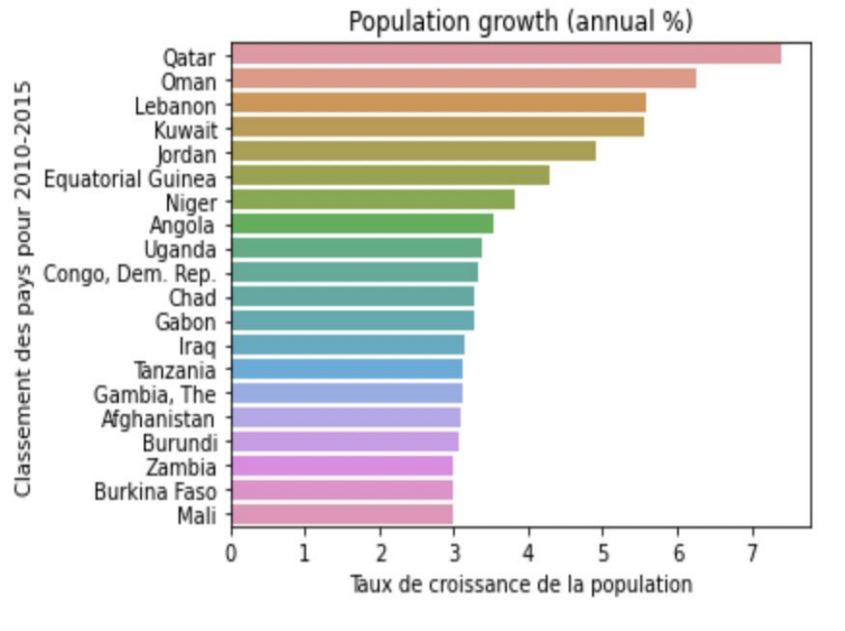


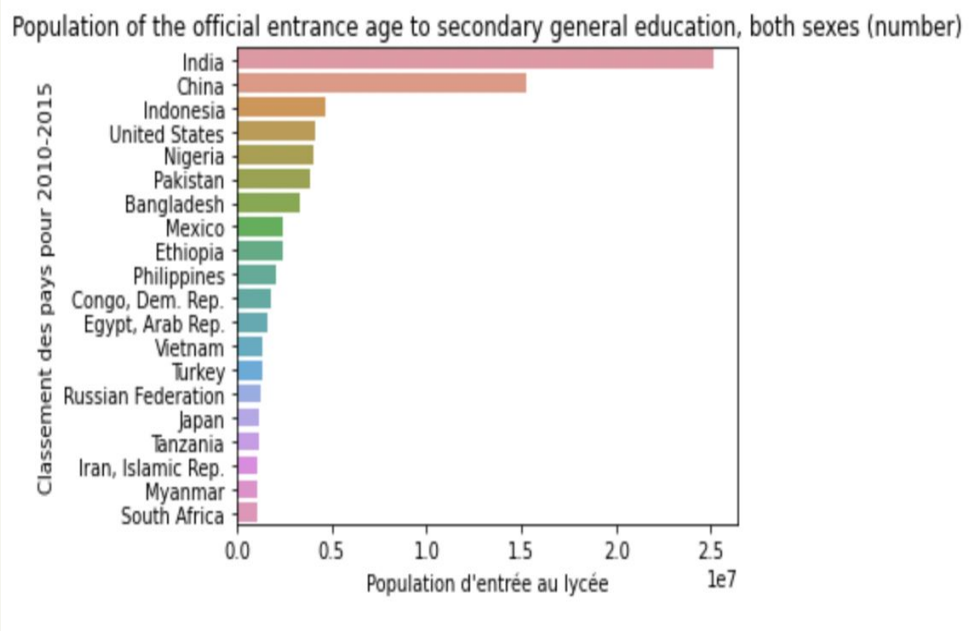
Démographie et éducation





Démographie et âge d'entrée dans le secondaire





Scoring

On utilise un classement par quartiles un score entre 0 et 3 pour chaque indicateur afin de noter le critère indicateur de chaque pays (3 pays pour la plus potentiel, 0 pour la moins potentiel).

Country Name	Average Years	Score_Internet_User	Score_Population_Growth	Score_Population_Secondary	Score_Population_15_24	Score_Enrolment_Tertiary	Score_
Afghanistan	7.483954e+06	0.0	3.0	3.0	3.0	0.0	
Albania	7.862124e+05	2.0	0.0	1.0	1.0	1.0	
Algeria	8.665805e+06	1.0	2.0	3.0	3.0	3.0	
American Samoa	-2.057934e-01	0.0	0.0	0.0	0.0	0.0	
Andorra	8.788385e+01	3.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	
Virgin Islands (U.S.)	4.243283e+01	2.0	0.0	0.0	0.0	0.0	
West Bank and Gaza	1.288042e+06	2.0	3.0	1.0	1.0	1.0	
Yemen, Rep.	6.307480e+06	1.0	3.0	3.0	3.0	0.0	
Zambia	3.252853e+06	1.0	3.0	2.0	2.0	0.0	
Zimbabwe	3.715917e+06	0.0	2.0	2.0	2.0	0.0	

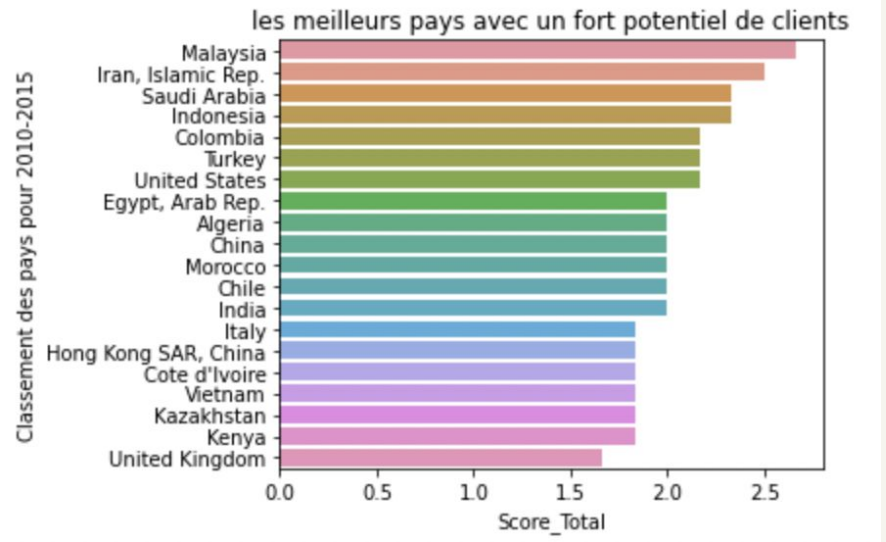
```
df3['Score_Internet_User'] = pd.qcut(df_Indict_Internet['Average Years'], 4, labels = False)
df3['Score_Population_Growth'] = pd.qcut(df_indic_Population_Growth['Average Years'], 4, labels = False)
df3['Score_Population_Secondary'] = pd.qcut(df_Population_Secondary['Average Years'], 4, labels = False)
df3['Score_Population_15_24'] = pd.qcut(df_Population_15_24['Average Years'], 4, labels = False)
df3['Score_Enrolment_Tertiary'] = pd.qcut(df_Enrolment_Tertiary['Average Years'], 4, labels = False)
df3['Score_Indic_Expenditure'] = pd.qcut(df_Indic_Expenditure['Average Years'], 4, labels = False)
```

Classement final

Sélectionner les pays pour lesquels tous les scores sont supérieures à une valeur (le 3ème quartile),

Par exemple un pays qui est dans le 3.0 sur plusieurs indicateurs aura un meilleur score qu'un pays qui n'y apparait qu'une fois (les meilleurs pays sont ceux pour lesquels la valeur de tous leurs indicateurs est supérieure au 3e quartile).

Donc pour réaliser ça nous ajoutons une colonne qui a fait une moyennes des score des indicateurs de chaque pays et après nous faisons un tri par cette valeur pour afficher les 30 pays avec un fort potentiel



Evolutions du potentiel

Méthode 1: Trouver les indicateurs qui ont été renseignés dans les années 2020 à 2100

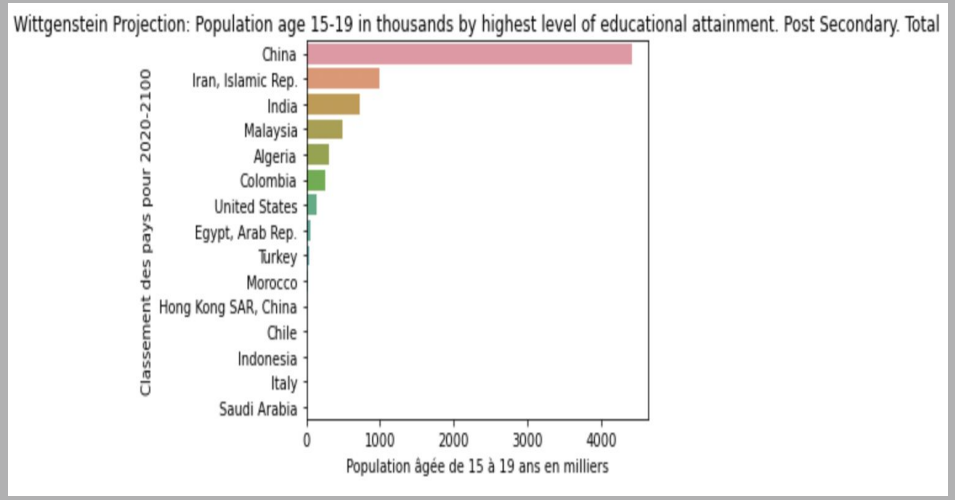
Méthode 2: Utiliser la librairie Scipy pour extrapoler des NAN

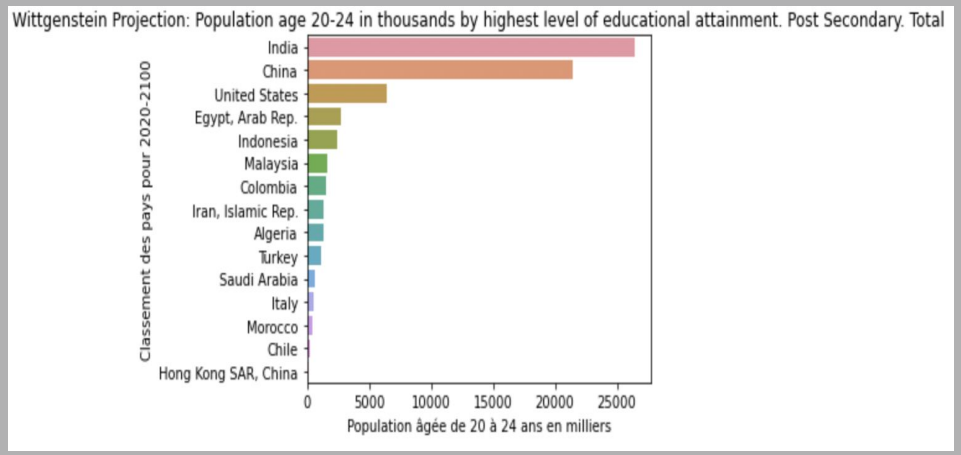
Résultat:
China', 'India', 'Malaysia', 'United States', 'Colombia','Egypt, Arab Rep.', 'Iran, Islamic Rep.', 'Algeria', 'Indonesia','Turkey', 'Hong Kong SAR, China', 'Morocco', 'Saudi Arabia', 'Chile','Italy'

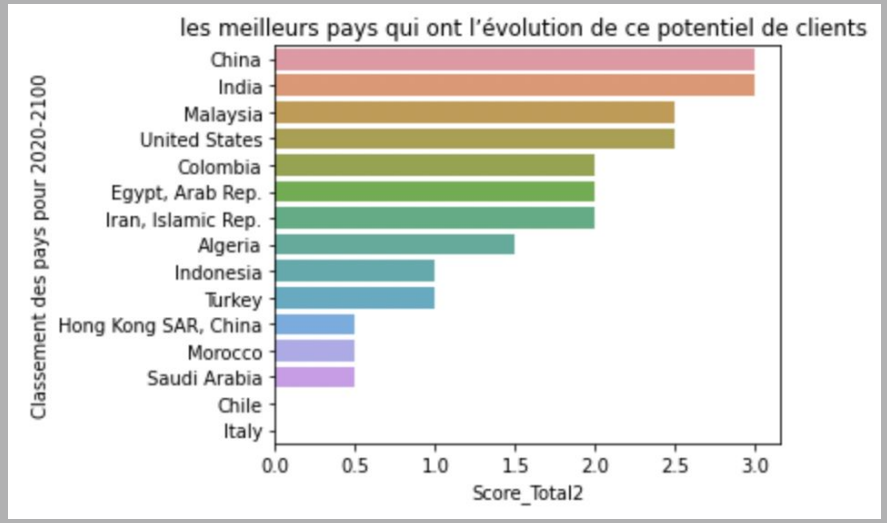
Méthode 1:

1. Trouver les indicateurs les mieux renseignés
2. Choisir 2 indicateurs :
    - Wittgenstein Projection: Population age 20-24 in thousands by highest level of educational attainment. Post Secondary. Total
    - Wittgenstein Projection: Population age 15-19 in thousands by highest level of educational attainment. Post Secondary. Total
3. Trouver les meilleurs pays pour chaque un des deux indicateur
4. Classement par quartiles un score entre 0 et 3 pour chaque indicateur
5. 10 top pays qui ont les meilleurs scores pour nos deux indicateurs

Scoring sur deux indicateurs prédictifs pour les années 2020-2100







Conclusion :

les pays commun dans les deux scoring précédents pour les années 2010_2015 et 2020 _2100

les pays qui ont eu un fort potentiel pour devenir des clients dans le passé et qui ont un fort potentiel pour devenir des clients à l'avenir

l'entreprise doit opérer en priorité dans 10 pays suivants:

- ★ United States
- ★ China
- ★ Iran, Islamic Rep.
- ★ Colombia
- ★ Algeria
- ★ Egypt, Arab Rep.
- ★ Indonesia
- ★ Turkey
- ★ Malaysia

Pistes d'amélioration

- Création d'un dashboard interactif
- Visualisation du classement sur une carte interactive
- etc

Thank you

@ Mitra.dadgar1367@gmail.com