

COURS EN LIGNE, NIVEAU LYCÉE ET
UNIVERSITÉ.

EXTENSION DU PÉRIMÈTRE
GÉOGRAPHIQUE.

FORMATION AUX TECHNOLOGIES: QUELLE CIBLE PAYS?



S O M M A I R E .

- Etude et chargement dataset.
- Analyse qualité dataset.
 - Par année: nombre valeurs manquantes.
 - Par pays, par année:
 - Taux de pourcentage valeurs non manquantes acceptables.
 - Filtrage dataset: élimination indicateurs avec taux qualité insuffisant.
- Etude typologie indicateurs.
 - Schéma de pertinence.
 - Sélection indicateurs.
- Complément dataset.
 - Import GDP, nettoyage.

SOMMAIRE (SUITE).

- Analyse qualité dataset complété.
 - Par pays, par année, par indicateur.
 - Sélection pays selon niveau qualité.
- Analyse performance par pays.
 - Calcul taux croissance par indicateur.
 - Liste pays fort potentiel.
- Estimation potentiel pays.
 - Liste pays par ordre décroissant.
- Pays prioritaires.

ETUDE ET CHARGEMENT DATASET.

Cinq dataset:

- Data: RangeIndex: 886930 entries, 0 to 886929 Data columns (total 70 columns).
- Country series: RangeIndex: 613 entries, 0 to 612 Data columns (total 4 columns).
- Country: RangeIndex: 241 entries, 0 to 240 Data columns (total 32 columns).
- Footnote: RangeIndex: 643638 entries, 0 to 643637 Data columns (total 5 columns).
- Series: RangeIndex: 613 entries, 0 to 612 Data columns (total 4 columns).

Dataset: data.

- Plusieurs lignes sont apparentées à des régions et non des pays.
- Constitution liste des régions, suppression lignes correspondantes.
- Après suppression: Int64Index: 798970 entries, 65970 to 886929 Data columns (total 70 columns).

Qualité data:

- df_data.duplicated().any() → False.
- df_data.isnull().any() → cf image.

Analyse des valeurs manquantes dataset.

Country Name	False
Country Code	False
Indicator Name	False
Indicator Code	False
1970	True
...	...
2085	True
2090	True
2095	True
2100	True
Unnamed: 69	True
Length: 70, dtype: bool	

ANALYSE QUALITÉ DATASET.

- Méthodologie d'analyse qualité.
 - Première étape:
 - Analyse par année: calcul nombre valeurs manquantes/nombre lignes total dataframe.
 - Fixation d'un seuil de qualité de données: 95%.
 - Comparaison calcul et valeur de seuil.
 - Conclusion ➔ Aucune année ne satisfait l'exigence de qualité.
 - Deuxième étape:
 - Nettoyage valeurs estimées (années 2021 à 2100).
 - Nouveau dataframe: 70 colonnes à 14.
 - Troisième étape:
 - Nouvelle étape de nettoyage de données: exclure les indicateurs avec niveau seuil qualité trop insuffisant.

Data columns (total 14 columns):			
#	Column	Non-Null Count	Dtype
0	Country Name	798970 non-null	object
1	Country Code	798970 non-null	object
2	Indicator Name	798970 non-null	object
3	Indicator Code	798970 non-null	object
4	2009	133538 non-null	float64
5	2010	232967 non-null	float64
6	2011	137275 non-null	float64
7	2012	138293 non-null	float64
8	2013	128642 non-null	float64
9	2014	105696 non-null	float64
10	2015	128534 non-null	float64
11	2016	15733 non-null	float64
12	2017	143 non-null	float64
13	2020	50820 non-null	float64

dtypes: float64(10), object(4)
memory usage: 91.4+ MB

ANALYSE QUALITÉ DATASET(SUITE).

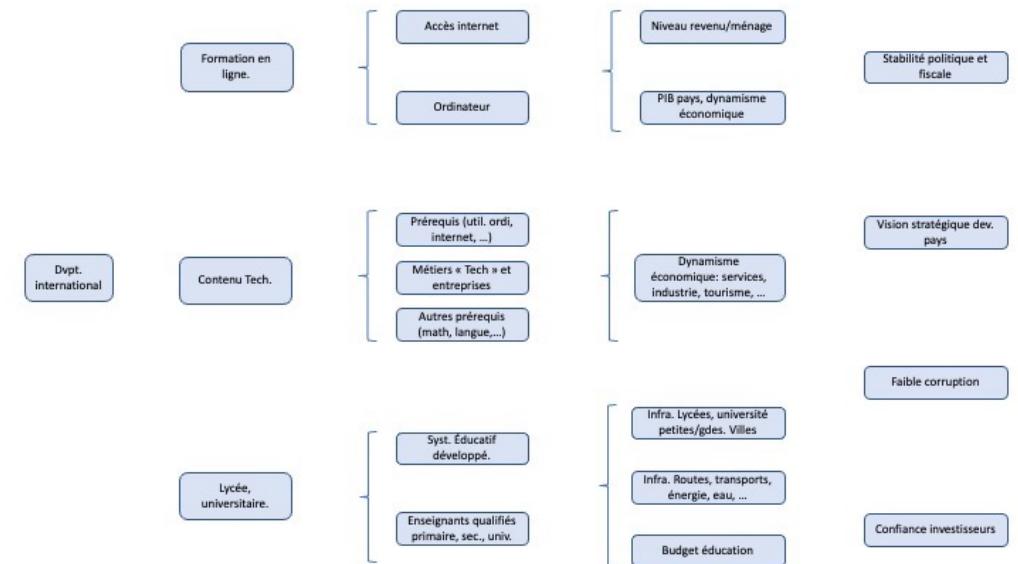
- **Méthodologie d'analyse qualité.**
 - Troisième étape (suite): détail analyse.
 - Pour chaque indicateur, on compte le nombre cellules pour lesquelles on a une valeur non manquante.
 - Calcul du rapport entre ce comptage et le nombre d'années (10).
 - Comparaison de ce rapport avec une valeur seuil déterminée.
 - La valeur seuil est volontairement relativement basse pour ne pas exclure tous les indicateurs, compte tenu de la qualité médiocre des données.
 - On exclue de l'analyse tous les indicateurs pour lesquels le rapport est en deçà de la valeur seuil.
 - Cette opération exclut 65% des données pour 2019. Le seuil est fixé à 60% (un indicateur ne doit pas comporter plus de 60% de valeurs manquantes).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 46916 entries, 67212 to 886613
Data columns (total 16 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   Country Name    46916 non-null  object  
 1   Country Code    46916 non-null  object  
 2   Indicator Name  46916 non-null  object  
 3   Indicator Code  46916 non-null  object  
 4   2009             46679 non-null  float64 
 5   2010             46786 non-null  float64 
 6   2011             46910 non-null  float64 
 7   2012             46905 non-null  float64 
 8   2013             46897 non-null  float64 
 9   2014             46892 non-null  float64 
 10  2015             46712 non-null  float64 
 11  2016             13651 non-null  float64 
 12  2017             0 non-null      float64 
 13  2020             0 non-null      float64 
 14  Total_non_manquants 46916 non-null  int64  
 15  Total_non_manquants_pc 46916 non-null  float64 
dtypes: float64(11), int64(1), object(4)
memory usage: 6.1+ MB
```

ETUDE TYPOLOGIE INDICATEURS.

- Les sous-indicateurs qui caractérisent le système éducatif appartiennent à plusieurs typologies d'indicateurs. Ces typologies sont recensées sur le lien [EdStats - Available Indicators](#).
- Il y a 19 typologies. Il est question ici de les passer en revue et de considérer leur intérêt dans la perspective de leur pertinence vis à vis des objectifs business:
 - Développement international
 - Métier entreprise:
 - Formation en ligne
 - Contenu orienté Tech
 - Public: secondaire et universitaire
 - Chargement dataset GDP pour compléter la liste d'indicateurs pertinents.

« Key drivers » pour identification kpi pertinents



COMPLÉMENT DATASET.

- Liste indicateurs pertinents: six indicateurs retenus:
 - ['gdp per capita (current us\$)', 'internet users (per 100 people)', 'labor force with advanced education (% of total)', 'teachers in secondary education, both sexes (number)', 'graduates from tertiary education, both sexes (number)', 'teachers in tertiary education programmes, both sexes (number)']
 - Après adjonction dataset GDP, intersection avec liste kpi, obtention nouveau dataframe avec caractéristiques suivantes:

Après intersection avec liste kpi.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 552 entries, 67215 to 884640
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Country Name     552 non-null    object  
 1   Country Code     552 non-null    object  
 2   Indicator Name   552 non-null    object  
 3   Indicator Code   552 non-null    object  
 4   2009              548 non-null    float64 
 5   2010              550 non-null    float64 
 6   2011              552 non-null    float64 
 7   2012              550 non-null    float64 
 8   2013              552 non-null    float64 
 9   2014              552 non-null    float64 
 10  2015              552 non-null    float64 
 11  2016              391 non-null    float64 
 12  2017              0 non-null     float64 
 13  2020              0 non-null     float64 
 14  Total_non_manquants 552 non-null int64  
 15  Total_non_manquants_pc 552 non-null float64 
dtypes: float64(11), int64(1), object(4)
memory usage: 73.3+ KB
```

ANALYSE QUALITÉ DATASET COMPLÉTÉ.

Méthodologie pour analyse qualité par pays, par indicateur:

Les deux fonctions suivantes permettent de filtrer davantage le dataset en augmentant davantage la qualité des données avec en contre partie une perte de pays potentiels.

La première fonction: assigne, pour un indicateur donné, la valeur 1 ou 0 selon que pour une année donnée, une valeur est présente ou non dans la cellule. A la fin de la lecture de chacune des cellules (indicateur-année), on fait la somme de la ligne. La somme est divisée par le nombre de cellules, on en déduit un rapport qui est comparée à une valeur seuil. Si le rapport est \geq seuil, on écrit la somme en regard de l'indicateur et on affecte 1 à la cellule 'Total'

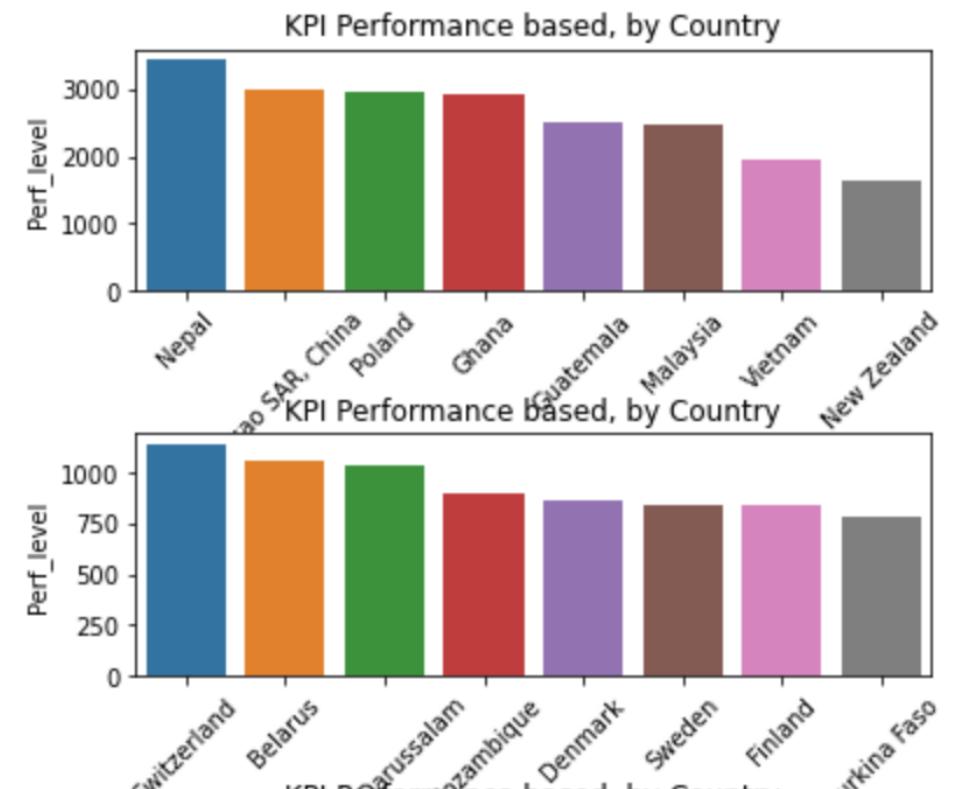
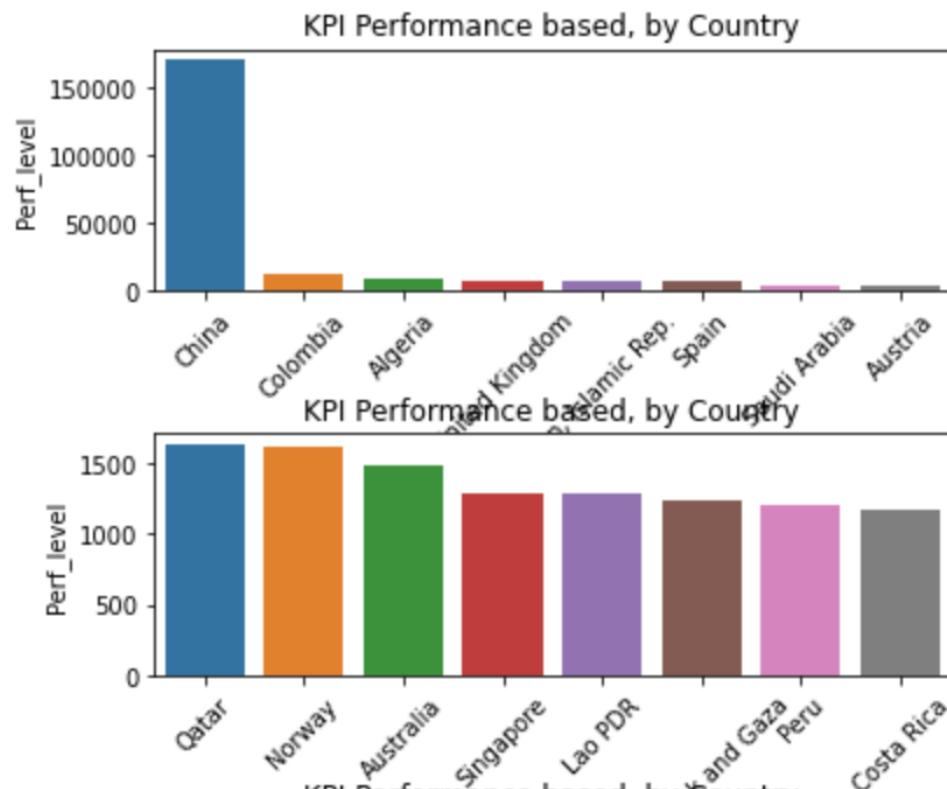
La deuxième fonction fait la somme, pour un pays donné, des valeurs de la colonne 'Total'. Cette somme est divisée par le nombre de kpi (6), et comparée à une valeur seuil. Si tous les indicateurs sont présents, on doit avoir une somme de 6. Si moins de 6 indicateurs sont présents on a une somme < 6 . En fonction de la valeur de la somme et de la valeur du seuil, on retient ou non le pays.

ANALYSE PERFORMANCE PAR PAYS.

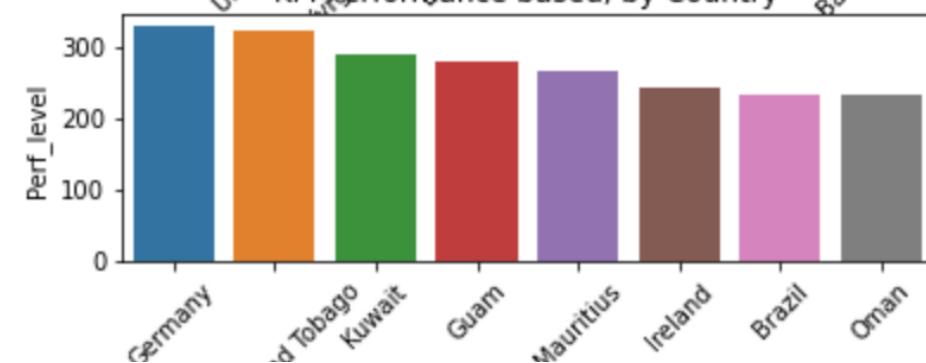
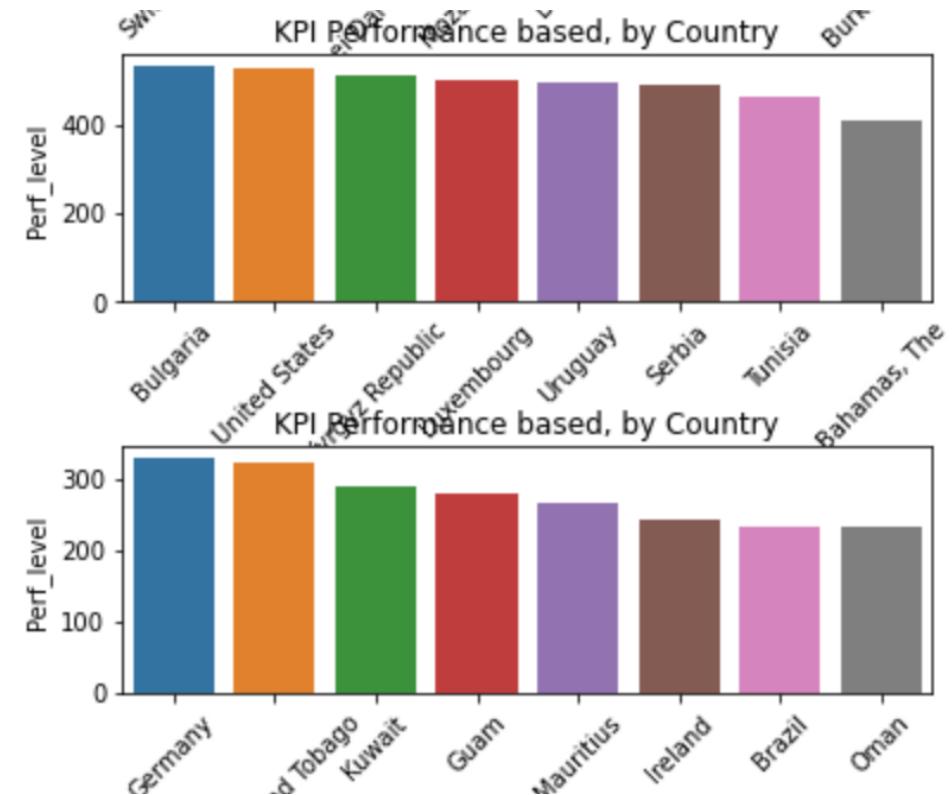
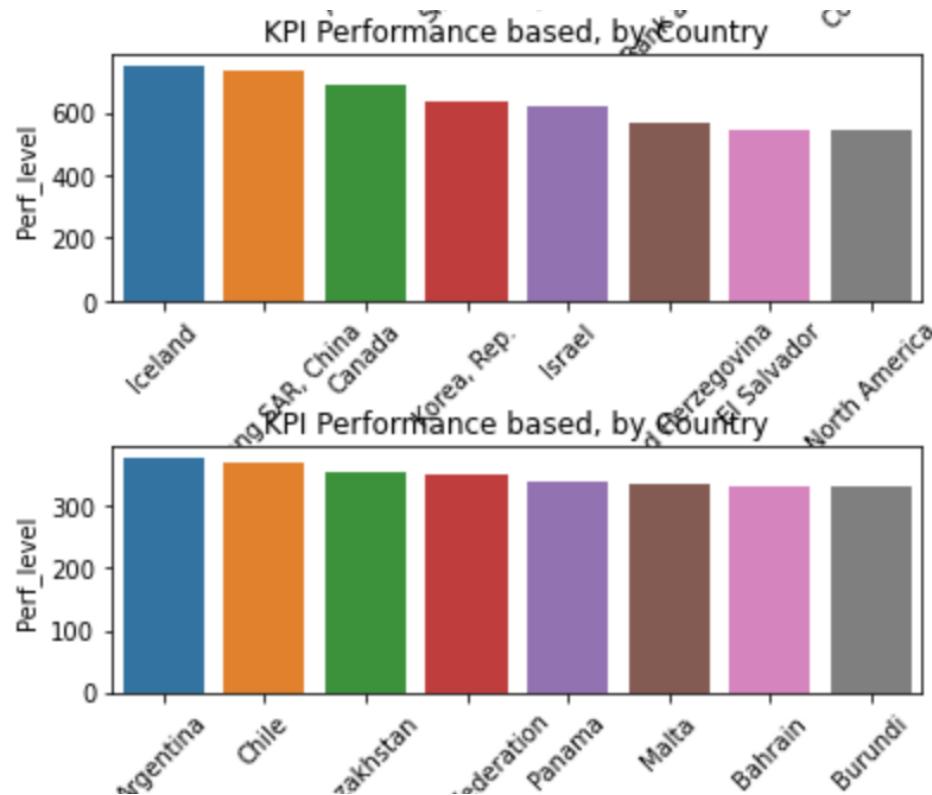
- Méthodologie:
 - Calcul valeur indicateur pour (année + 1) – année.
 - Sommer toutes les valeurs et diviser par le nombre d'éléments pour obtenir une moyenne.
 - Obtention d'une liste de 177 pays à fort potentiel.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 177 entries, 35 to 132
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Country Name  177 non-null    object  
 1   Perf_level    177 non-null    float64 
dtypes: float64(1), object(1)
memory usage: 4.1+ KB
```

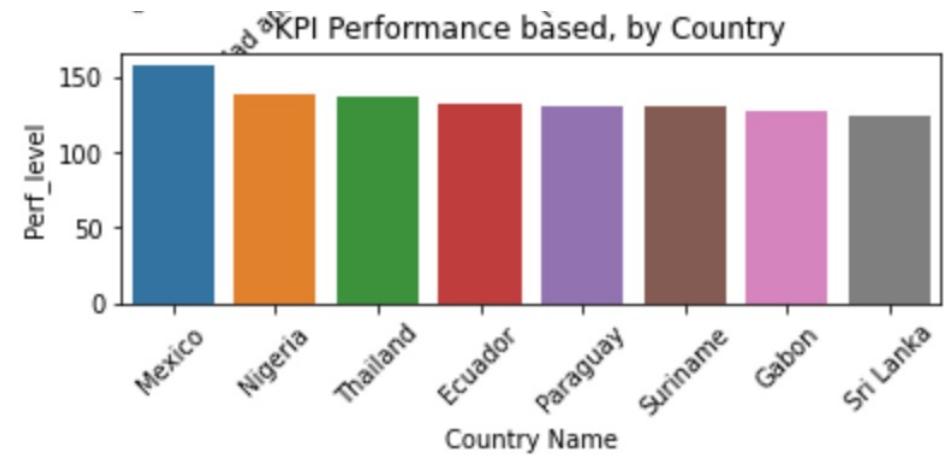
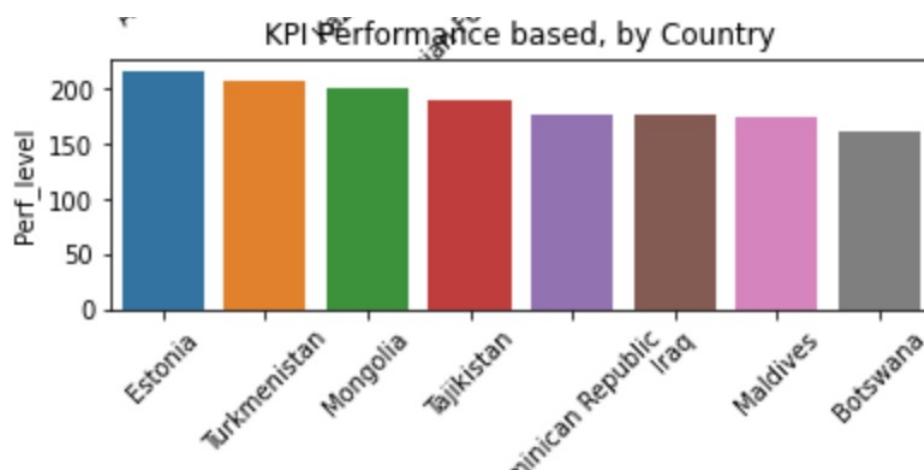
ANALYSE PERFORMANCE PAR PAYS.



ANALYSE PERFORMANCE PAR PAYS.



ANALYSE PERFORMANCE PAR PAYS.



ESTIMATION POTENTIEL PAYS.

- Méthodologie d'estimation du potentiel des pays.
- On se base sur la performance moyenne des pays. Cette valeur moyenne est celle qui est reprise comme performance future.
- Compte tenu du biais d'estimation, on ne procède que sur une période relativement courte de trois ans.

PAYS PRIORITAIRES.

- Conclusion:**

- Les données sont à l'origine de mauvaise qualité.
- A cause de ce niveau de qualité faible, certains pays sont disqualifiés car trop peu d'indicateurs présents à cause d'une mauvaise qualité.
- Trop peu d'indicateurs induisent une relative faible performance.
- Certains pays sont mal classés à cause de cette faible performance. C'est le cas du Brésil qui aurait probablement mérité d'être parmi les pays prioritaires.

	Country Name	Country Code	Indicator Name	Indicator Code	Future_perf_lev
0	China	CHN	graduates from tertiary education, both sexes ...	SE.TER.GRAD	172875.671429
1	Colombia	COL	graduates from tertiary education, both sexes ...	SE.TER.GRAD	11503.814286
2	Algeria	DZA	graduates from tertiary education, both sexes ...	SE.TER.GRAD	7858.771429
3	Spain	ESP	graduates from tertiary education, both sexes ...	SE.TER.GRAD	7670.757143
4	Iran, Islamic Rep.	IRN	teachers in tertiary education programmes, bot...	SE.TER.TCHR	7657.471429
5	United Kingdom	GBR	graduates from tertiary education, both sexes ...	SE.TER.GRAD	7184.471429
6	Nepal	NPL	teachers in secondary education, both sexes (n...	SE.SEC.TCHR	3401.2125
7	Macao SAR, China	MAC	gdp per capita (current us\$)	NY.GDP.PCAP.CD	3022.556512
8	Ghana	GHA	teachers in secondary education, both sexes (n...	SE.SEC.TCHR	2877.55
9	Malaysia	MYS	teachers in secondary education, both sexes (n...	SE.SEC.TCHR	2559.771429