PRESENTATION

ACADEMY

SOMMAIRE

- Problématique liée à l’entreprise
- Validation de la qualité des données
- Description du Dataset principal
- Sélection des indicateurs via les informations pertinentes
- Description des indicateurs retenus
- Statistiques des indicateurs retenus
- Top 10 des pays sélectionnés (Score synthétique)

PROFIL ACADEMY

Qui somme nous ?

L'entreprise est une organisation dont l'objectif est d'étendre ses activités à l'international (niveau lycée et études supérieurs).

Notre mission

Notre mission est d'utiliser l'analyse des données sur l'éducation pour informer et faciliter l'expansion internationale de l'entreprise, en identifiant des opportunités de marché prometteuses.

Notre vision

Notre vision est de devenir un leader mondial dans les services éducatifs en ligne en exploitant intelligemment les données pour améliorer l'accès à l'éducation à l'échelle mondiale.

Notre objectif

Notre objectif est de trouver des pays attractifs afin de contribuer à l'amélioration de l'accès à l'éducation à travers le monde.

Quels sont les pays à fort potentiel pour nos services d'éducation en ligne ? Comment les départager ?

PROBLÉMATIQUE

THE WORLD BANK
IBRD • IDA

PRÉ-ANALYSE DES DONNÉES DU PORTAIL EDUCATION STATISTICS (EDSTATS)

- 🔎Indicateurs pertinents.
- 📊Indicateurs statistiques
- 🎯Top 10 pays

VALIDATION DE LA QUALITÉ DES DONNÉES

~~EdStatsCountry-Series.csv~~

Les jeux de données EdstatsCountry-Series et EdstatsFootNote sont exclus de notre analyse en raison de leur moindre pertinence et de leur apport limité à nos objectifs d'analyse.

~~EdStatsFootNote.csv~~

EdStatsData.csv

se compose de
- 886 930 lignes
- 70 colonnes.

Le taux moyen de complétion des données est de 13,9%.
Pas de doublons, evolution des indicateurs par pays (1970 - 2100)
c'est notre jeu de données principal

EdStatsCountry.csv

se compose de
- 241 lignes
- 32 colonnes

Pas de doublons.
Informations géographiques, données économique.

EdStatsSeries.csv

se compose de
- 3665 lignes
- 21 colonnes

Il présente une valeur intéressante à savoir les Topic (thème de l'indicateur)

VALIDATION DE LA QUALITÉ DES DONNÉES

NOTRE JEU DE DONNÉE PRINCIPAL

- Il est essentiel de souligner que les valeurs manquantes sont entre 0 et 95 % dans ce dataset

- Le data set est composé comme tel :

Une ligne représente un indicateur pour son pays par année (1970 - 2100)

- Informations utiles : une plage temporelle pour connaitre l'évolution dans le temps, sélectionner des indicateurs pertinents.

NOTRE JEU DE DONNÉE PRINCIPAL

- Nous remarquons un schéma récurrent d'évolution des données tous les 5 ans depuis 1970 jusqu'en 2010.
- La dernière année exploitable est 2015.
- Les années 2020 à 2100 sont des années de prédiction.
- les donnée de 2016 à 2020 sont inexploitables
- Les années 2000 et 2005 sont les années avec le plus de données significatives et la plus fournie avec environ 80% de taux de remplissage.

NOTRE JEU DE DONNÉE PRINCIPAL

Préconisations :

- Supprimer la colonne Unnamed: 69 en raison de l'absence de données exploitables.
- Prendre en compte que la période de 2020 à 2100 correspond à des données predicitves.
- Exclure la période de 2016 à 2019 en raison du manque de données exploitables.
- Évaluer la nécessité de nettoyer ou de supprimer la colonne Country Name en fonction des besoins spécifiques de l'analyse.

SELECTIONS DES INFORMATIONS PERTINENTES

```python
# Liste des indicateurs restants
indicateurs_restants = indicator_list_deleted

# Liste des indicateurs à filtrer
# Indicateurs potentiels que nous souhaitons sélectionner
indicators_potential = ['GDP', 'age 15-19', 'age 20-24', 'Population, ages 15-24, total',
                        'Population of the official age for secondary education, both sexes (numbe
                        'Population of the official age for tertiary education, both sexes (number
                        'computers', 'Gross enrolment ratio, secondary, both sexes (%)',
                        'Gross enrolment ratio, secondary, gender parity index (GPI)',
                        'Gross enrolment ratio, tertiary, both sexes (%)',
                        'Gross enrolment ratio, tertiary, gender parity index (GPI)',
                        'DHS: Secondary completion rate',
                        'Enrolment in tertiary education per 100,000 inhabitants, both sexes',
                        'Internet users']

# Filtrer les indicateurs qui sont dans la liste indicators_potential
indicators_pertinents = [indicator for indicator in indicateurs_restants if any(ind in indicator f

# Créer un DataFrame avec les indicateurs pertinents
data_potential_indicators = data_EdStats[data_EdStats['Indicator Name'].isin(indicators_pertinents

print(len(data_potential_indicators['Indicator Name'].unique().tolist()))
data_potential_indicators['Indicator Name'].unique().tolist()
```

```
68
```

Step 2
Recherche d'indicateurs pertinents et compréhension des termes pour les indicateurs : https://datatopics.worldbank.org/education/indicators

Nb : 68

Step 4
Analyser le taux de remplissage avec pour seuil le quartile 3 = 50% .

Nb : 35

```python
# Liste de tout les indicateurs
indicator=data_EdStats['Indicator Name'].unique()
nb_indicator=len(indicator)
country_=data_EdStats['Country Name'].unique()
nb_country=len(country_)


print('Nombre d indicateurs : ')
print(nb_indicator)
print('Nombre de pays : ')
print(nb_country)

nb_country*nb_indicator==data_EdStats.shape[0]
indicator.tolist()
```

```
Nombre d indicateurs :
3665
Nombre de pays :
242
```

1

Nb : 3665

Step 1
Brainstorming
- Population visée
- Aspects économique
- Education : taux de scolarisation, % d'élèves diplômés..

2

3

Nb : 35

Step 3
Analyse par plage temporelle et par thème (Topic)

4

5

Nb : 8

Step 5
Affiner la selection avec la recherche par mot clefs.

```python
# Liste des mots-clés pour la filtration (en minuscules)
mots_cles = ['gdp',
             'internet',
             'computer',
             'population, ages 15-24, total',
             'completed secondary',
             'completed tertiary']

# Filtrer les indicateurs en utilisant une compréhension de liste
indicators_filtered = [indicator for indicator in list_indicators_refined if any(mot in indicator.

# Afficher la liste des indicateurs filtrés
print(indicators_filtered)
print(len(indicators_filtered))
```

```
['Barro-Lee: Percentage of population age 15-19 with secondary schooling. Completed Secondary', 'Barr
o-Lee: Percentage of population age 15-19 with tertiary schooling. Completed Tertiary', 'Barro-Lee: P
ercentage of population age 20-24 with secondary schooling. Completed Secondary', 'Barro-Lee: Percent
age of population age 20-24 with tertiary schooling. Completed Tertiary', 'GDP per capita (current US
$)', 'Internet users (per 100 people)', 'Personal computers (per 100 people)', 'Population, ages 15-2
4, total']
8
```

```python
# Liste des topics souhaités
topics_souhaites = ["Attainment",
                    "Education Equality",
                    "Tertiary",
                    "Economic Policy & Debt: National accounts: US$ at current p
                    "Secondary", "Population", "Infrastructure: Communications"]

# Filtrer le DataFrame pour ne conserver que les lignes avec les topics souhaite
data_academy = data_academy[data_academy['topic'].isin(topics_souhaites)]

# Obtenir la liste des noms des indicateurs correspondants
noms_des_indicateurs = data_academy['indicator_name'].unique()

# Afficher la liste des noms des indicateurs correspondants
print(len(noms_des_indicateurs))
print(noms_des_indicateurs)
```

```
35
```

LES INDICATEURS RETENUS

- Educations
- Economie
- Numérique
- Démographique

BAR.SEC.CMPT.1519.ZS

Pourcentage de la population âgée de 15 à 19 ans ayant suivi un enseignement secondaire.

BAR.TER.CMPT.1519.ZS

Pourcentage de la population âgée de 15 à 19 ans ayant suivi un enseignement supérieur.

BAR.SEC.CMPT.2024.ZS

Pourcentage de la population âgée de 20 à 24 ans ayant suivi un enseignement secondaire.

BAR.TER.CMPT.2024.ZS

Pourcentage de la population âgée de 20 à 24 ans ayant fait des études supérieures.

NY.GDP.PCAP.CD

PIB par habitant (USD courants)

IT.NET.USER.P2

Utilisateurs d'Internet (pour 100 personnes)

IT.CMP.PCMP.P2

Ordinateurs personnels (pour 100 personnes)

SP.POP.1524.TO.UN

Population âgée de 15 à 24 ans, total

- Le taux de personnes fréquentant l'enseignement supérieur varie davantage d'un pays à l'autre que le taux de personnes allant au lycée.
- Pourles indicateurs numérique qui sont exprimés en pourcentage, il y a trois groupes de pays principaux. Le premier groupe se situe autour de 15%, le deuxième autour de 45%, et le troisième autour de 70%.

STATISTIQUE DES INDICATEURS RETENUS

BAR.SEC.CMPT.1519.ZS

BAR.TER.CMPT.1519.ZS

BAR.SEC.CMPT.2024.ZS

BAR.TER.CMPT.2024.ZS

NY.GDP.PCAP.CD

IT.NET.USER.P2

IT.CMP.PCMP.P2

SP.POP.1524.TO.UN

Pour certains indicateurs liés à la population, à l'économie et à l'éducation au niveau universitaire, il y a des pays qui se distinguent car ils ont des valeurs très différentes par rapport à la plupart des autres pays. On les appelle des "outliers".

STATISTIQUE DES INDICATEURS RETENUS

BAR.SEC.CMPT.1519.ZS

BAR.TER.CMPT.1519.ZS

BAR.SEC.CMPT.2024.ZS

BAR.TER.CMPT.2024.ZS

NY.GDP.PCAP.CD

IT.NET.USER.P2

IT.CMP.PCMP.P2

SP.POP.1524.TO.UN

STATISTIQUE DES INDICATEURS RETENUS

BAR.SEC.CMPT.1519.ZS

BAR.TER.CMPT.1519.ZS

BAR.SEC.CMPT.2024.ZS

BAR.TER.CMPT.2024.ZS

NY.GDP.PCAP.CD

IT.NET.USER.P2

IT.CMP.PCMP.P2

SP.POP.1524.TO.UN

TOP 10 DES PAYS RETENUS

(SCORE SYNTHÉTIQUE AVEC PONDÉRATION)

Avec Pondération (Moyenne Pondérée des Indicateurs) :

Nous pouvons créer un score synthétique en combinant les scores des différents indicateurs pour obtenir un classement global. Un moyen courant de faire cela est d'utiliser une moyenne pondérée des scores des indicateurs, où nous attribuons des poids à chaque indicateur en fonction de son importance relative:

1) Assignez des poids à chaque indicateur en fonction de son importance. Par exemple, si nous estimons que l'indicateur "NY.GDP.PCAP.CD" (PIB par habitant) est plus important que l'indicateur "IT.NET.USER.P2" (utilisateurs d'Internet), nous pouvons attribuer un poids plus élevé à l'indicateur PIB.

2) Calculer le score synthétique pour chaque pays en combinant les scores normalisés de chaque indicateur en utilisant les poids attribués. Nous pouvons utiliser une moyenne pondérée pour cela. (Pour chaque indicateur, nous calculons un score normalisé en utilisant la méthode de normalisation. La normalisation consiste à mettre à l'échelle les valeurs de l'indicateur de manière à ce qu'elles varient de 0 à 1. Cela se fait en soustrayant la valeur minimale de l'indicateur à chaque valeur, puis en divisant par la plage (la valeur maximale moins la valeur minimale). Le résultat est un score compris entre 0 et 1.)

3) Trier les pays en fonction de leur score synthétique pour obtenir un classement global.

TOP 10 DES PAYS RETENUS

(SCORE SYNTHÉTIQUE AVEC PONDÉRATION)

0.5
0.4
0.4
0.4

EVOLUTION DANS LE TEMPS (INDICATEURS PAR PAYS)

MERCI

Des question ?

Academy