Vous êtes Data Scientist dans une start-up, qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

Votre lead Data Scientist vous confie une première mission d'analyse exploratoire, pour déterminer si les données sur l'éducation de la banque mondiale permettent d'informer le projet d'expansion.

Voici les différentes questions auxquels vous devriez répondre durant votre analyse:

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

Pour une meilleur reussite du projet, votre lead vous suggère les étapes ci dessous:

A) Rappel de la problématique
- Comprendre la problématique
- Choix des indicateurs

B) Présentation du jeu de données
- Importer les librairies
- Importer les données
- Comprendre les données

C) Filtrage des données selon des critères de sélection
- Pré-filtrage des données
- Identifier les Nan
- Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?
- Identifier les duplicatas
- Suppression des colonnes inutiles ou entièrement vides
- Sélectionner les indicateurs principals et supprimer les autres
- Suppression des pays trop petits (Population < 3M d'habitants)
- Suppression des pays très pauvres (PIB trop bas)
- Suppression des anciennes années et mal renseignées
- Suppression des années qui ont 50% de NaN
- Sélectionner les années en fonction des indicateurs

C) Choix de 10 meilleurs pays
- Scoring par pays
- Graphes

D) Choix des meilleurs zones géographiques
- Scoring par région
- Graphes

E) Choix des pays où l'entreprise doit opérer en priorité

Par ailleurs, trouvez ci dessous les indicateurs et les colonnes

population = ['SP.POP.TOTL']
nb_eleves = ['SP.POP.1215.TO.UN','SP.POP.1524.TO.UN','SP.POP.AG25.TO.UN']
internet = ['IT.NET.USER.P2']
computer = ["IT.CMP.PCMP.P2"]
PIB = ['NY.GDP.PCAP.PP.CD']
inscription université = ['SE.TER.ENRL']
inscription lycée = ['UIS.E.3']
depense_pub_education = ['SE.XPD.TOTL.GD.ZS']

AIDARA Chamsedine

aidarachamsedine10@gmail.com