Projet n°2 : “Analyse de données de Système Educatifs”

Soutenance de Jury - Projet n°2 - Octobre 2021

Ali Naama

Sommaire

I. Description de la problématique et péparation du jeu de données

Description de la problématique

Description :

academy

L'entreprise « Academy" cherche à s'étendre à l'international. Pour se faire, elle nous sollicite pour mener une analyse de données afin de trouver des axes de ciblages stratégiques par pays.

- Academy intervient dans le domaine de la EdTech (Educational Technology : technologies de l'éducation).
- Academy est une start-up qui fournit des formations en ligne pour des niveaux lycée et université.
- Les données d'analyse sont fournies par la banque mondiale rubrique « EdStats »

BANQUE MONDIALE

Objectifs :

Analyser les données de la banque mondiale afin de déterminer quels sont les pays avec les plus forts potentiels actuel et futur d'expansion de la startup Academy pour trouver dans quels pays opérer en priorité.

❑ Dans ce projet, nous ferons des recommandations stratégiques à partir de données de systèmes éducatifs issues de la World Bank.

Présentation du jeu de données

Informations sur les variables :

World Bank. (2013). Education statistics. Retrieved June 6, 2020, from http://data.worldbank.org/data-catalog/ed-stats

La requête sur tous les indicateurs EdStats de la Banque mondiale contient plus de 4 000 indicateurs comparables

au niveau international qui décrivent l'accès à l'éducation, la progression, l'achèvement, l'alphabétisation, les enseignants, la population et les dépenses. Les indicateurs couvrent le cycle d'éducation de l'enseignement primaire à l'enseignement professionnel et supérieur.

- Nombre de fichiers : 5
- Nombre de variables : 4000

II. Préparation du jeu de données

Présentation du jeu de données

BANQUE MONDIALE

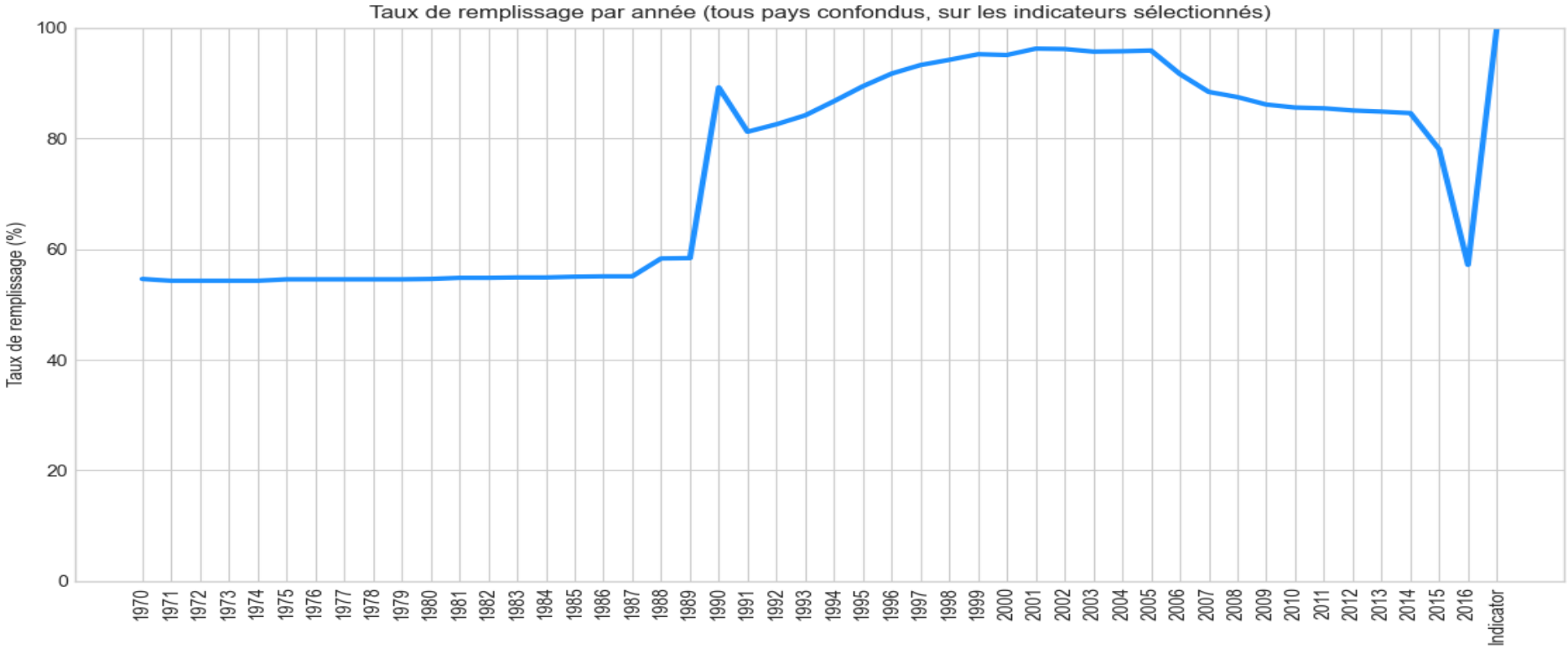
EdStatsCountry.csv	**Informations globales sur l'économie de chaque pays du monde (et de zones géographiques)** Taille : 241 lignes (1 par pays / zone) , 32 colonnes Quelques valeurs manquantes Aucun doublon
EdStatsCountry-Series.csv	**Informations sur la source des données contenues dans EdStatsCountry** Taille : 613 lignes, 4 colonnes Pas de valeur manquante (sauf Unnamed : 3" qui est une colonne uniquement composée de NaN) Aucun doublon
EdStatsData.csv	**Donne l'évolution de nombreux indicateurs pour tous les pays et certains groupes de pays** Taille : 886 930 lignes, 70 colonnes données depuis 1970 **Nombreuses valeurs manquantes** Aucun doublon
EdStatsFootNote.csv	Contient des Informations sur l'année d'origine des données et les incertitudes sur les données) Taille : 643 638 lignes, 4 colonnes Pas de valeur manquante (sauf Unnamed : 4 qui est une colonne uniquement composée de NaN) Aucun doublon
EdStatsSeries.csv	Informations sur les indicateurs socio économiques disponibles dans EdStatsData. Taille : 3665 lignes, 21 colonnes 6 colonnes vides pour lesquelles il manque toutes les valeurs. Il manque plus de 80 % des données dans 10 autres colonnes de la table Aucun doublon

Présentation du jeu de données

Fichier	Nb lignes	Nb colonnes	Taux remplissage moyen	Doublons	Description
EdStatsCountry.csv	241	32	69.5%	0	Liste des pays avec leurs données principales
EdStatsSeries.csv	3665	21	28.3%	0	Liste des indicateurs avec description, unité, période, etc...
EdStatsData.csv	886930	70	13.9%	0	Données de chaque indicateur par pays et par année
EdStatsCountry-Series.csv	613	4	75.0%	0	Description des différentes séries de données (majoritairement provenance)
EdStatsFootNote.csv	643638	5	80.0%	0	Commentaire pour chaque couple série de données / pays

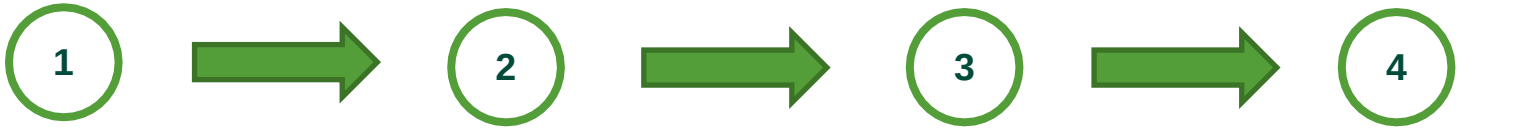
- On notera que le fichier « **EdStatsData.csv** », contient le plus grand nombre de lignes mais contient aussi le taux moyen de remplissage le plus bas

Présentation du jeu de données



- On notera que l'on peut se concentrer sur la plage de temps : **2005 à 2015** ou le taux de remplissage global est de 80% minimum. Nous nous concentrerons donc sur cette plage de temps afin de limiter les données à manipuler et nous concentrer sur les données les plus complètes

Stratégie d’analyse



Découvrir les données

Quelles informattions avons-nous ?

Quels catégories d’indicateurs ciblés ?

Quel niveau de Qualité ?

Quelle période ciblée ?

Identifier les indicateurs clés

Quantités de données manquantes ?

Comparer les Régions et Pays

Quels indicateurs choisir pour les comparaisons ?

Analyse des résultats obtenus

Quels sont les régions du monde et pays à cibler par Academy?

Quel est le potentiel pour chaque pays?

Comment identifier le potentiel des pays choisis?

Outils utilisés pour l'analyse

Nom	Utilisation	Fonctions spécifiques
Anaconda 1.7.2	Gestion de package Gestion d'environnement virtuel	Conda : installation de package via le terminal
Jupyter Notebook 6.1.4	Structurer la démarche Executer code pas à pas Expliquer la démarche (markdown)	
PyCharm 2021.1	Test et développement	IDE Community Edition, Debug, Synchro Git
Python 3.9.5	Moteur Python Gestionnaire de librairies	Moteur d'exécution
Pandas 1.2.4	Librairie de manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation, pivotage, autre
Matplotlib 3.4.2 **Seaborn 0.11.2** **Numpy 1.20.3**	Génération de graphiques Gestion des densités de probabilité	Barplot, Scatterplot, lineplot, distplot, heatmap Calcul statistique

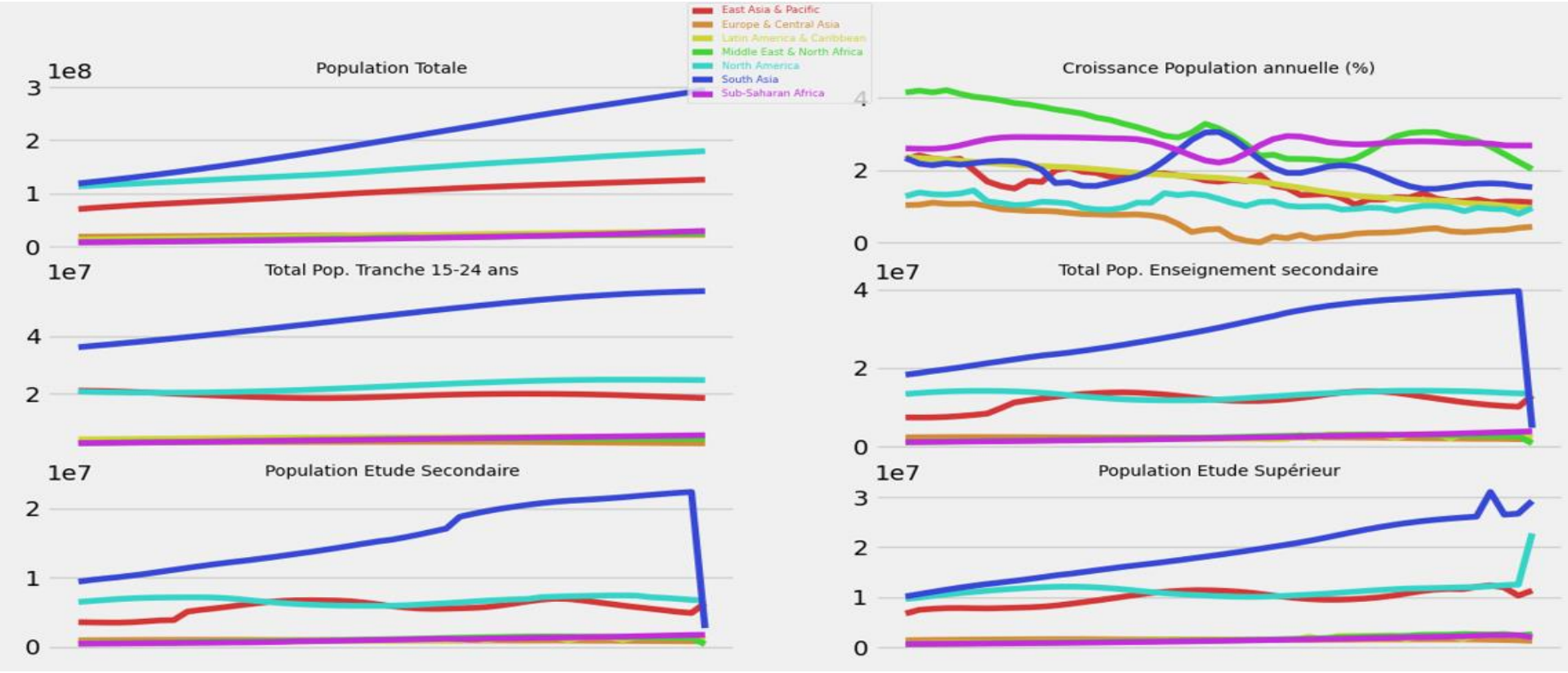
III. Analyse Exploratoire

Observations – Indicateurs Clés

Ayant listé et ciblé la majorité des indicateurs clés en fonction de 4 catégories, nous identifions les indicateurs suivants (ficher : « EdStatsSeries.csv »)

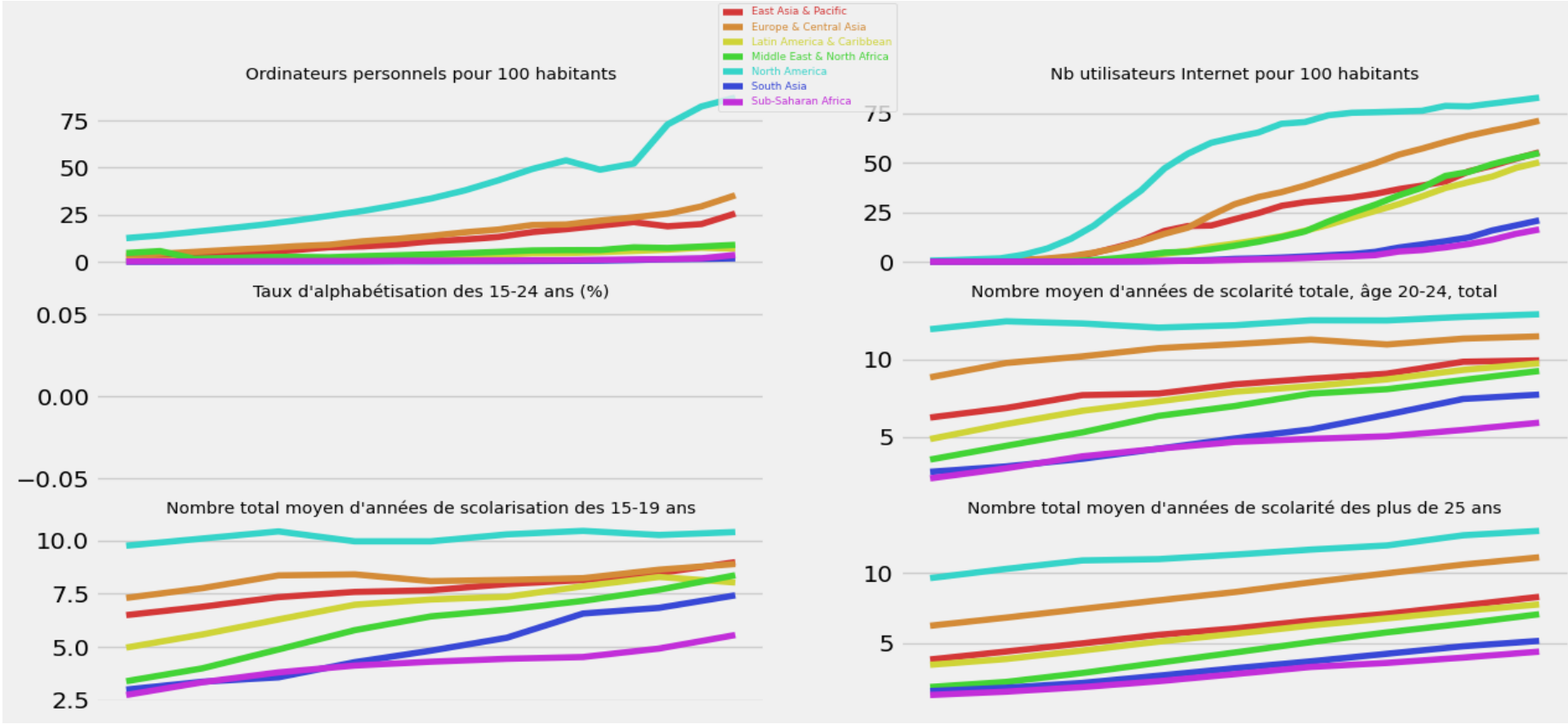
#	Indicateurs Clés	Catégorie (Topic)	Description
#1	SP.POP.TOTL-Population, total SP.POP.GROW-Population growth (annual %) SP.POP.1524.TO.UN-Population, ages 15-24, total SP.SEC.TOTL.IN-Population of the official age for secondary education, both sexes (number) SP.SEC.UTOT.IN-Population of the official age for upper secondary education, both sexes (number) SP.TER.TOTL.IN-Population of the official age for tertiary education, both sexes (number)	**Démographie et Scolarisation**	Potentiel de croissance de la population. Chercher à dénombre la population lycéens et universitaires, équivalent à la tranche d'âge 15-24 ans.
#4	SE.ADT.1524.LT.ZS-Youth literacy rate, population 15-24 years, both sexes (%) BAR.SCHL.2024-Barro-Lee: Average years of total schooling, age 20-24, total BAR.SCHL.1519-Barro-Lee: Average years of total schooling, age 15-19, total BAR.SCHL.25UP-Barro-Lee: Average years of total schooling, age 25+, total	**Niveau d'éducation**	Taux d'Alphabétisation et de scolarisation par tranche d'âge
#3	NY.GDP.MKTP.PP.CD-GDP, PPP (current international $) NY.GDP.PCAP.PP.CD-GDP per capita, PPP (current international $)	**Economique**	Evolution du PIB / PNB
#2	IT.CMP.PCMP.P2-Personal computers (per 100 people) IT.NET.USER.P2-Internet users (per 100 people)	**Accès Internet**	Possession d'un PC et Internet pour 100 habitants

Observations – Indicateurs Economiques par Région



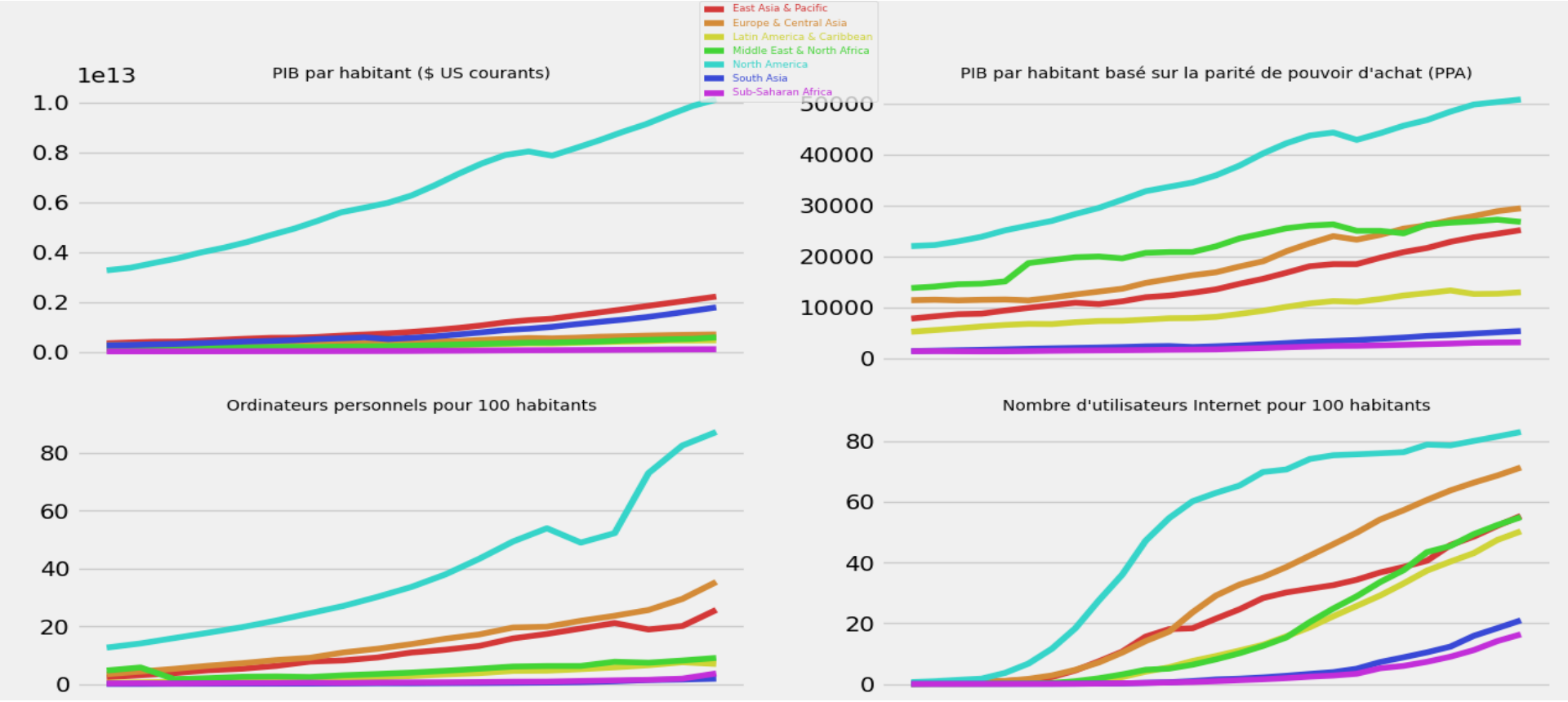
- En terme de croissance de population, on constate que les zones : East-Asia, North America, South Asia sont très dynamique

Observations – Indicateurs du niveau d'éducation par Région



- On constate que le nombre d'utilisateurs Internet pour 100 habitants est très importants dans les zones : North America, Europe, Middle East, Latine América, East Asia

Observations – Indicateurs du niveau d'accès à Internet par Région

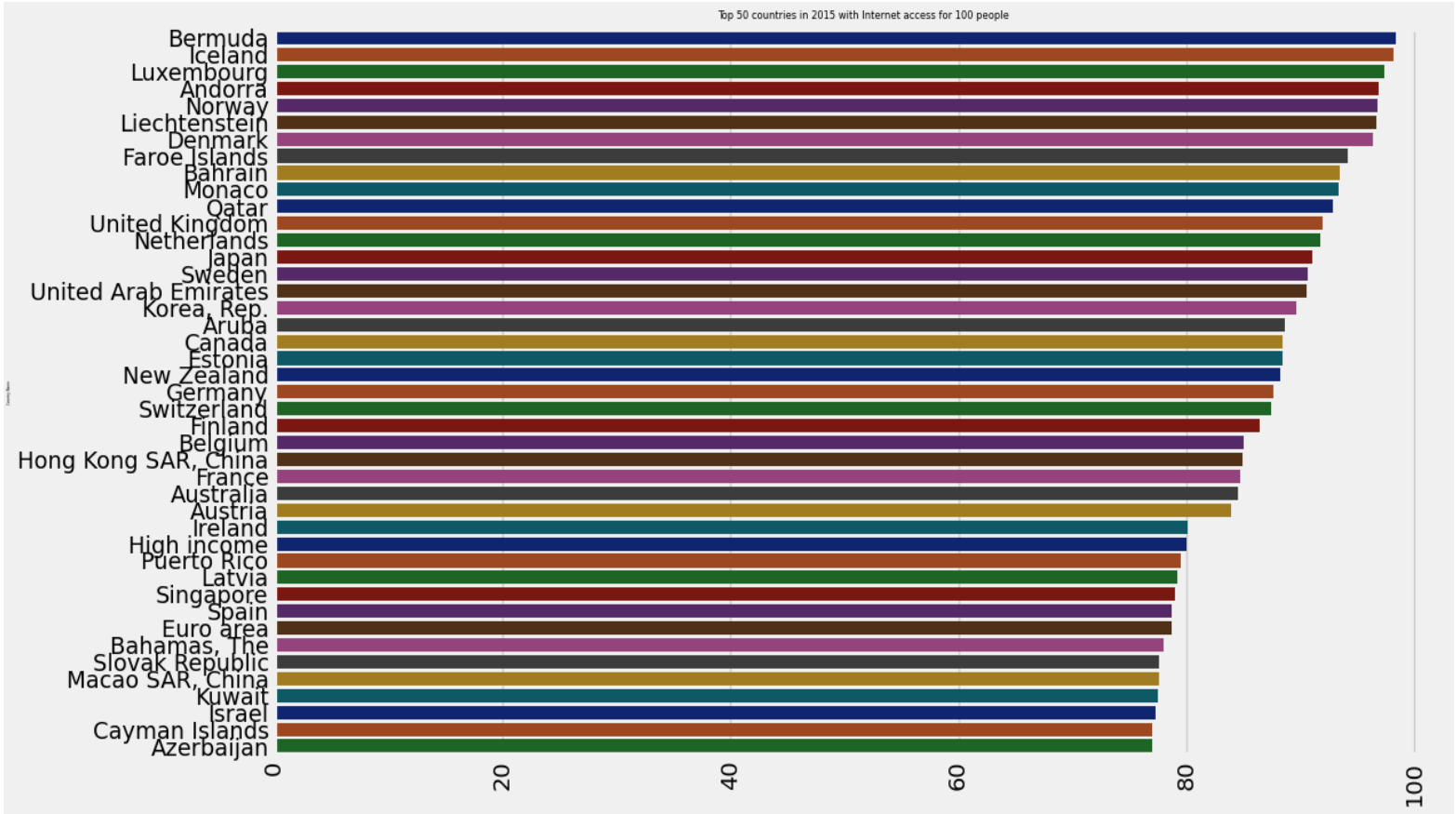


- On constate que les PIB (PPA) est très importants dans les zones : North America, Europe, Middle East, East Asia

IV. Constats

Analyse

❑ Top 50 des pays avec Accès internet pour 100 habitants



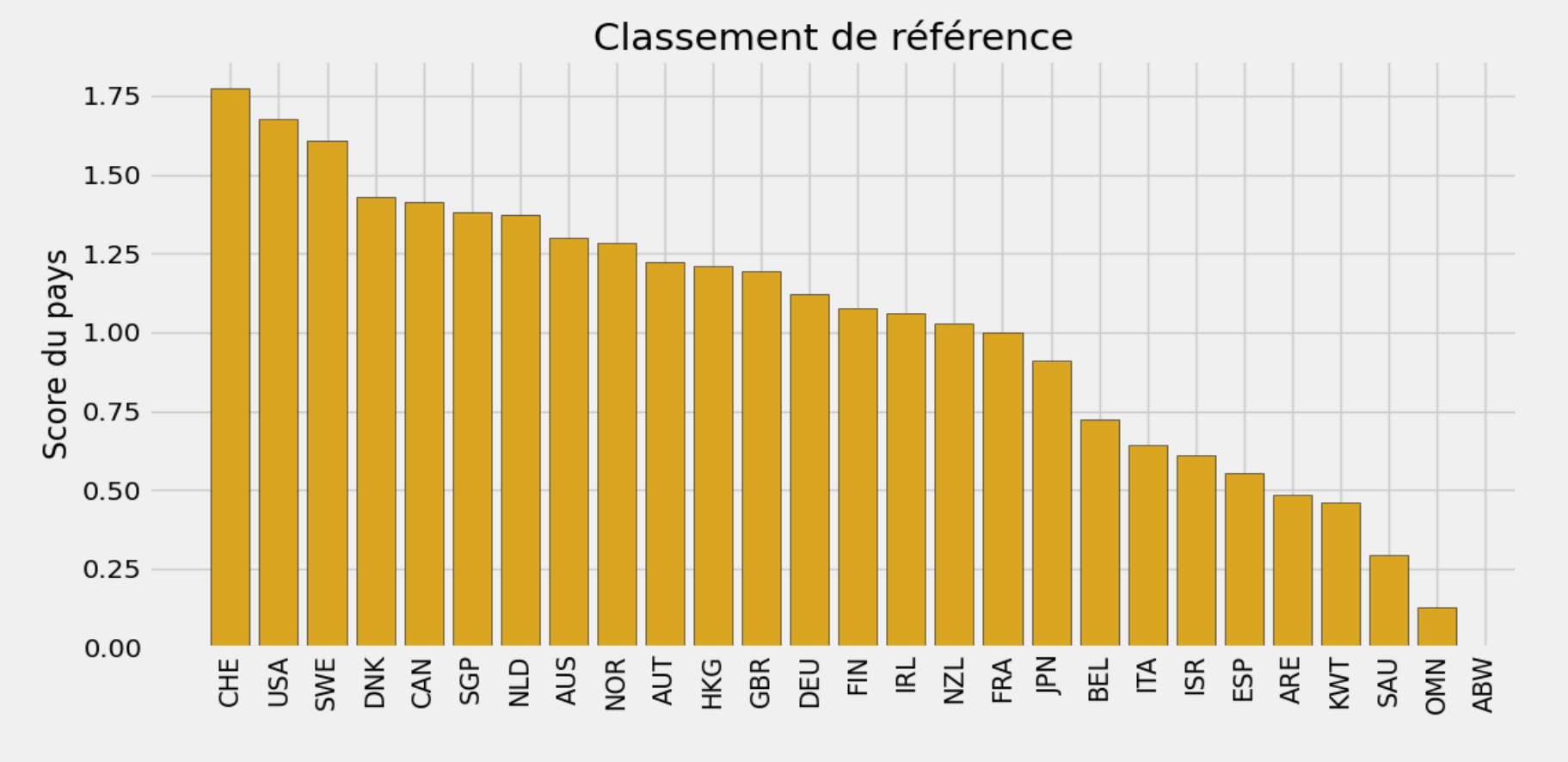
Constats

- Les analyses précédentes sur les différents indicateurs nous donne de grandes tendances par régions.
- Afin d'affiner nos résultats et pouvoir obtenir un indicateur synthétique nous allons opter pour la création d'un indicateur quantitatif nommé « score » qui va cibler directement un pays et porter une synthèse quantitative des indicateurs clés suivants :
  - IT.CMP.PCMP.P2 : Ordinateurs personnels (pour 100 habitants)
  - NY.GDP.PCAP.PP.CD : Pouvoir d'achat
  - SP.POP.1524.TO.UN : Total Population des 15-24 ans,
  - SP.POP.GROW : Santé : Dynamique de la croissance démographique (% annuel)
  - IT.NET.USER.P2 : Utilisateurs d'Internet (pour 100 habitants)
- Le principe est le suivant : On utilise une fonction spécifique
  - On divise chaque valeur d'un indicateur par le maximum de sa colonne de façon à avoir des valeurs entre 0 et 1.
  - On va ensuite calculer la somme pondérée des colonnes de chaque indicateur retenu pour avoir le score du pays.
  - On applique une pondération similaire à tous les indicateurs en première itération
  - On restitue le résultat sous forme de graphique à bar par pays pour afficher le score calculé

```
# For each row of the dataframe, we calculate the score - Add all column together
for country_code, row in scoring_data.iterrows():
    score = 0
    for column, coef in coefficients.items():
        score += row[column] * coef
    scoring_data.at[country_code, score_column_name] = score
```

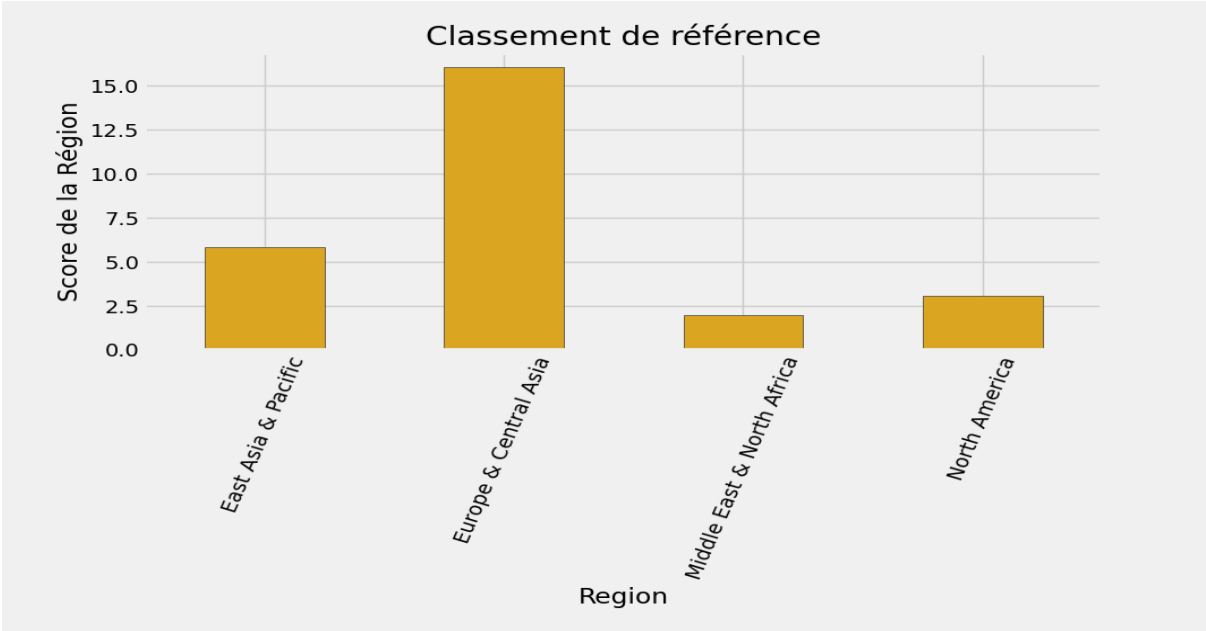
V. Résultats et recommandations

Scoring de référence par Pays



Résultats et recommandations

- ❑ Le plus pertinent serait de cibler les 20 premiers pays du classement par région géographique
- ❑ Mise en place d'un score agrégé par région pour permettre de cibler les régions les plus dynamiques
- ❑ Il faut cibler en priorités les Régions suivantes :
  - ❑ Europe & Central Asia
  - ❑ East Asia & Pacific
  - ❑ North America
  - ❑ Middle East



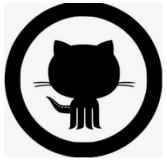
VI. Conclusion

Conclusion

- Ce projet m'a permis de manipuler plusieurs fichiers csv et d'en extraire les informations pertinentes pour orienter notre client vers les régions/pays les plus pertinents

- Le point à retenir est :
  - La sélection des variables pertinentes
  - La possibilité de tracer des sous graphes avec matplotlib
  - Le « feature engineering » de variable pour le calcul du score avec une fonction de somme
  - Agréger des données pour les restituer sous forme graphique afin de synthétiser les informations clés nécessaire à la prise de décision

Git



Merci de votre attention