Data scientist
Projet 2

Analysez des données de systèmes éducatifs

Dolores Valide

Problématique

- Domaine : Elearning
- Projet : expansion à l'international
- Public cible : niveau lycée et université
- Mission : analyse exploratoire des données
- Objectif : déterminer les pays à fort potentiel de clients, leur évolution dans le temps. Proposer une liste de pays prioritaires.

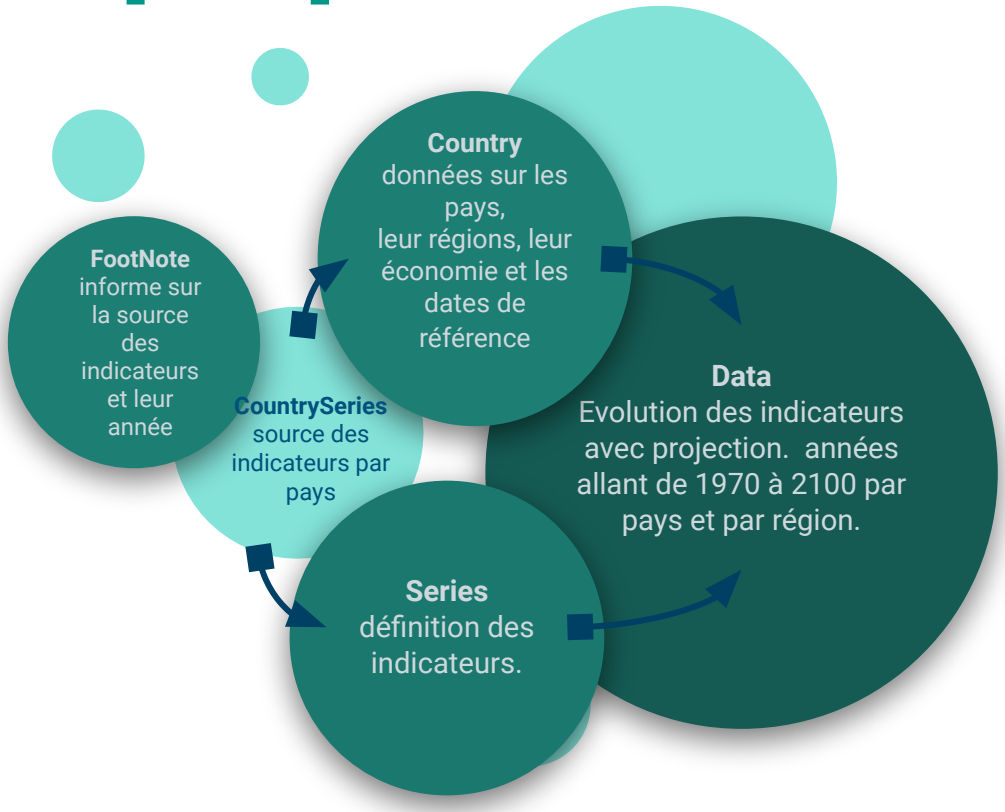
Sommaire

1. Découverte des données
2. Exploration des données
3. Filtrage et analyse
4. Conclusions

1.
Découverte des données

- Le jeu de données et les liens entres fichiers
- Compréhension des variables
- Qualité du jeu données

Des datasets qui dépendent les uns des autres.



Quelle est la qualité de ce jeu de données?

EdStatFootNote : 643 638 lignes pour 5 colonnes. 0 valeurs manquantes.

EdStatData : 886 930 lignes 70 colonnes , années 1970 - 2100. 4 premières colonnes sans aucunes valeurs manquantes. La dernière entièrement vide et les 65 autres contiennent environ 80% de données manquantes.

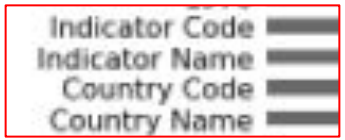
CountrySeries : 613 lignes pour 4 colonnes. 0 valeurs manquantes. 1 colonne entièrement vide.

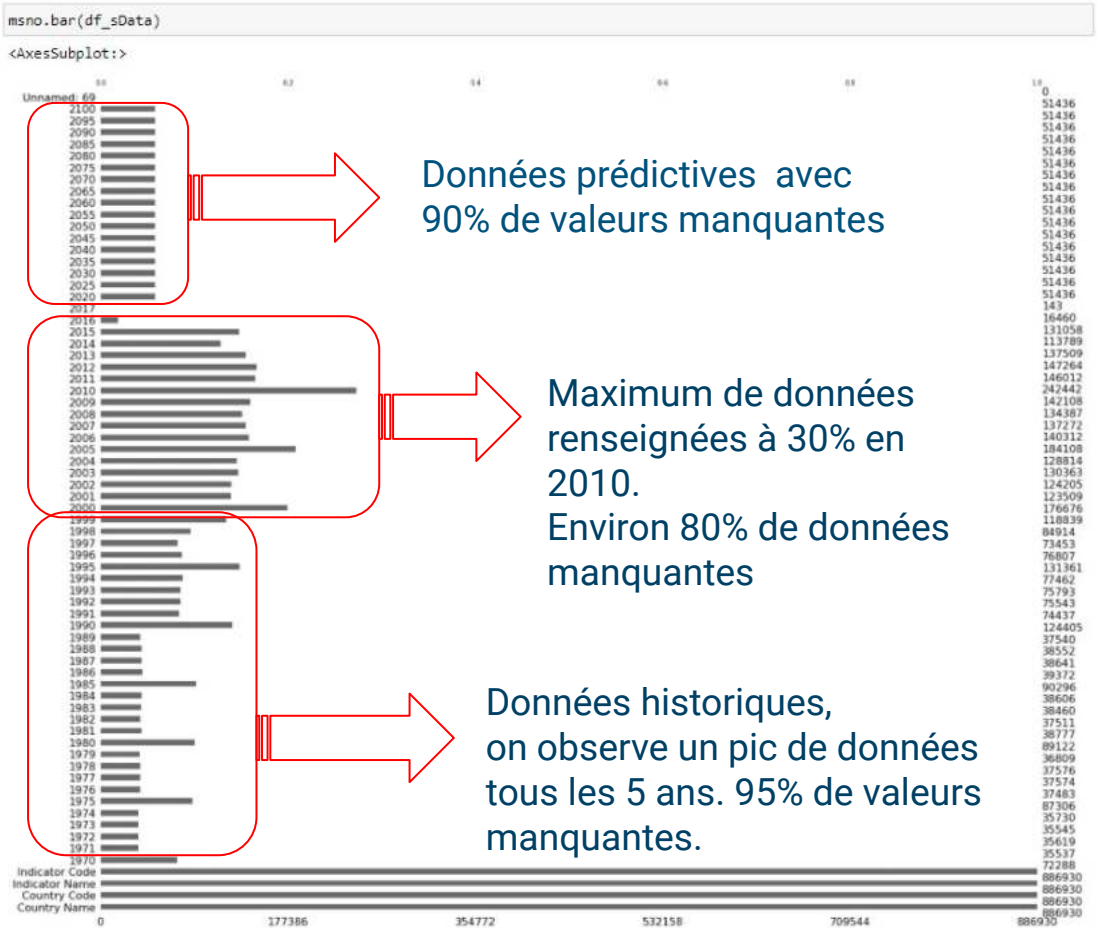
0 doublons pour tous les fichiers

EdStatSeries : 3665 lignes et 21 colonnes, 1 colonne inutile et 6 totalement vides. 5 colonnes totalement pleines, 60 à 80 % de valeurs manquantes sur les autres

EdStatCountry : 241 lignes pour 32 colonnes , environ 30 % de données manquantes.

Analyse du remplissage pour sData





Données prédictives avec
90% de valeurs manquantes

Maximum de données
renseignées à 30% en
2010.
Environ 80% de données
manquantes

Données historiques,
on observe un pic de données
tous les 5 ans. 95% de valeurs
manquantes.

2.
Exploration des données

- Exploration de notre dataset
- Répartition des données

Combien de pays et de régions sont renseignés?

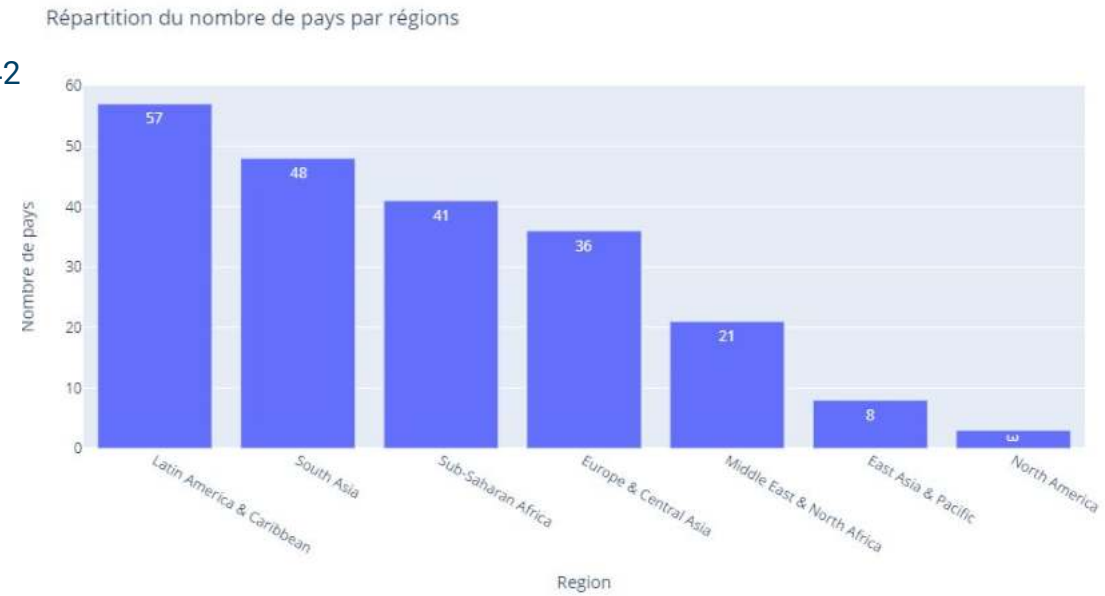
Nombres de pays et zones confondues : 242

Nombres de pays selon l'ONU : 197

Nombre de régions : 7

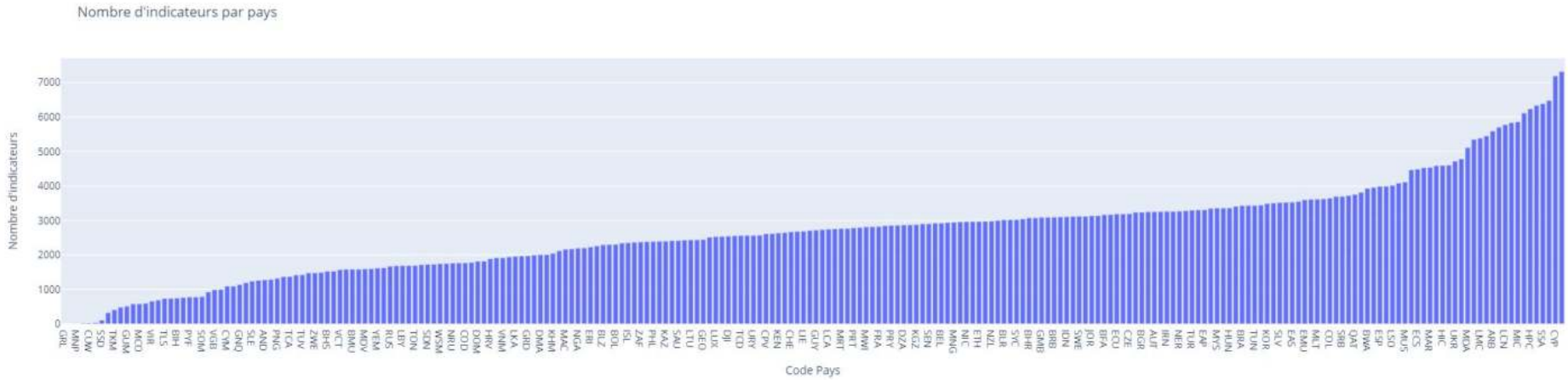
Nombre de groupes économiques : 5

Nombre d'indicateurs : 3665



Graphique obtenu à partir de sCountry

Il y a une disparité dans le nombre d'indicateurs différents renseignés par pays

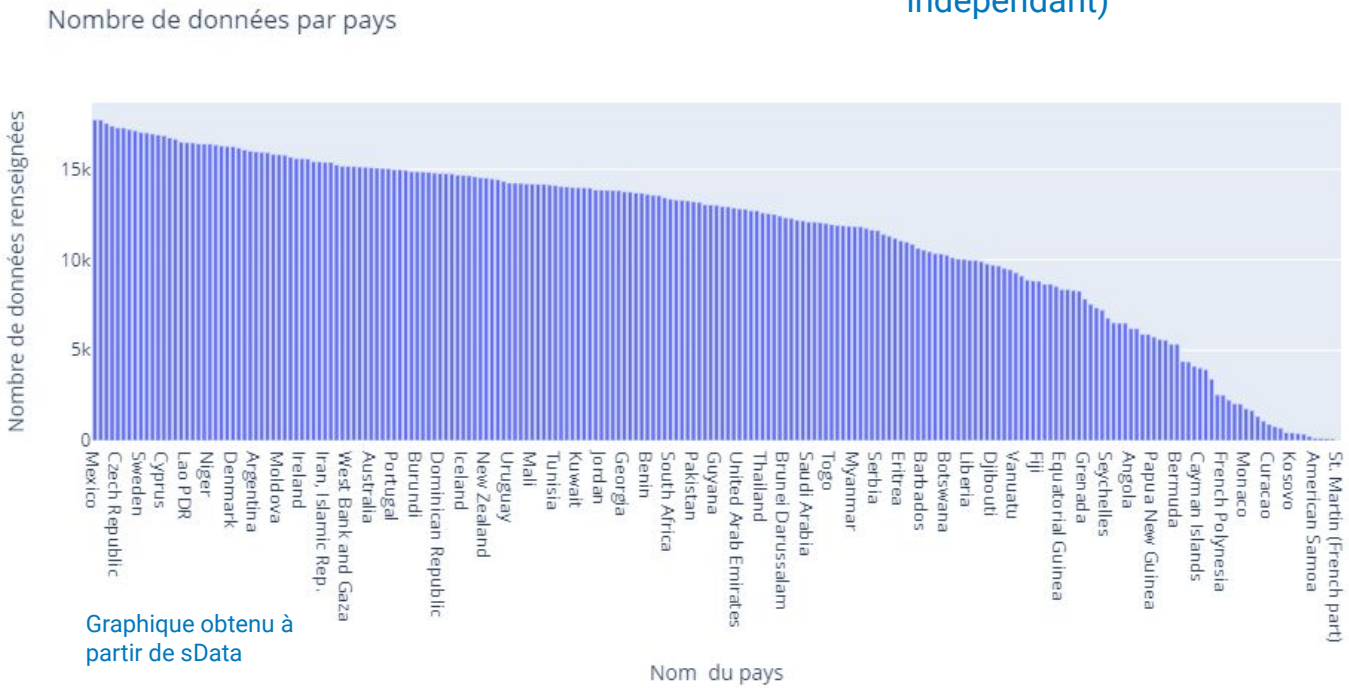


Les "pays" ayant le plus d'indicateurs renseignés sont en fait les regroupements de pays (économique ou continental)

Graphique obtenu à partir de Foot_Series qui est la combinaison de FootNote et CountrySeries

Il y a une inégalité dans la répartition des données dû à la population

- Les pays qui ont le moins de données renseignées sont les petits pays comme les îles ou les parties de pays(partie française de St Martin) .
- On retrouve aussi les pays jeunes (nouvellement indépendant)



Graphique obtenu à partir de sData

1er filtre :
Les pays à très faible population peuvent être écartés

Il y a une inégalité de renseignement des données en fonction des années

- La période la mieux renseignée est de 1999 - 2015
- Les données prédictives sont quasi nulles

2eme filtre : Nous allons nous cantonner à cette période pour notre analyse.

Graphique obtenu à partir de FootNote

3.
Filtrage et analyse

- Choix des indicateurs
- Choix des valeurs
- Statistiques descriptives par régions
- Affinage de la liste des pays
- Pays cibles pour l’expansion
- Évolution de ces pays

academy

Parmis les 4000 indicateurs voici ceux que nous avons sélectionnés

Démographique

Population cible

Lycéens : Total d'élèves public et privé inscrit au lycée quelques soit leur âge et leur sexe.
Étudiants : Total d'étudiants inscrit dans le tertiaire public ou privé.
15-24 ans : population totale des 15-24 ans

- UIS.E.3
- SE.TER.ENRL
- SP.POP.1524.TO.UN

Numérique

Materiel et installation informatique

Accès à internet : Taux de pénétration d'internet vu sur les 3 dernier mois.
Accès à un ordinateur personnel : Taux de pénétration d'ordinateurs destiné à un usage privé.

- IT.NET.USER.P2
- IT.CMP.PCMP.P2

Economique

Moyens financiers

Investissement du gouvernement dans l' éducation : total des dépenses exprimées en pourcentage du PIB

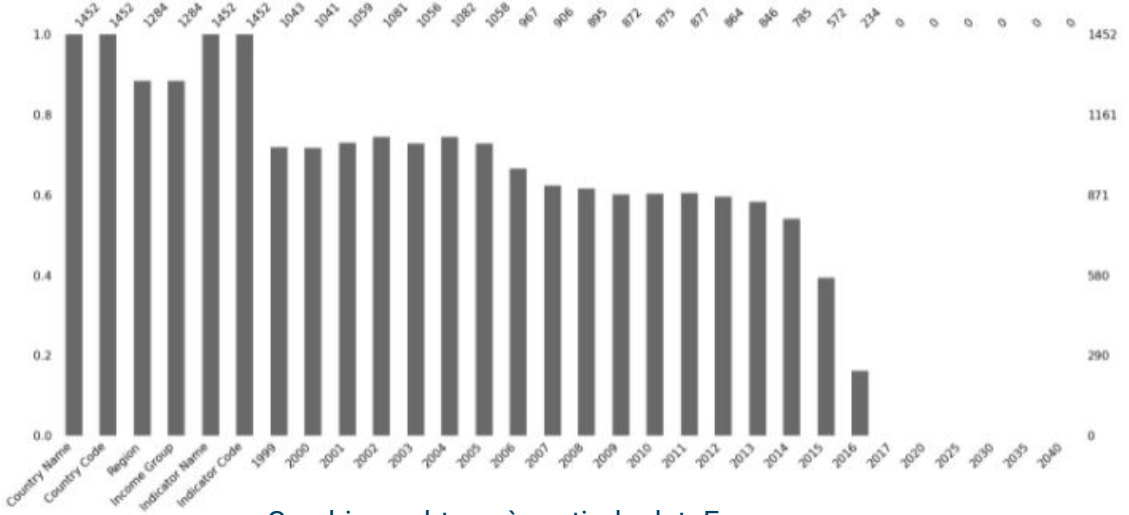
- SE.XPD.TOTL.GD.ZS

Comment est le remplissage du dataFrame avec les indicateurs sélectionnés?

- On avoisine les 40% de données manquantes de 2007 à 2014
- Il nous reste 1452 lignes pour 30 colonnes

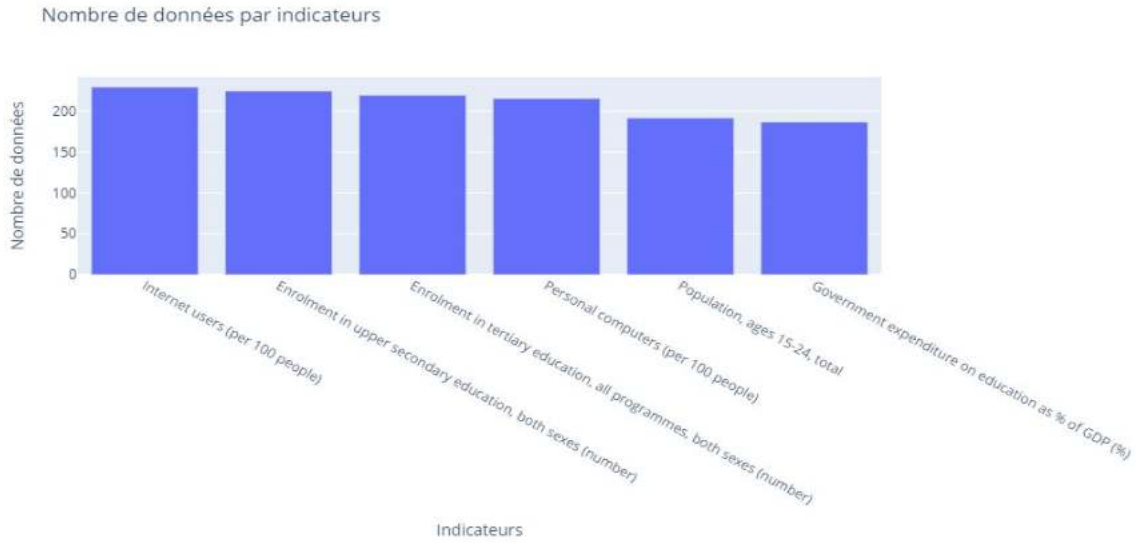
Maintenant nous allons récupérer la dernière valeure non nulle pour chaque pays et chaque indicateur puis préparer notre jeu de données pour l'analyse

```
Entrée [68]: msno.bar(filtre_indicateurs)
Out[68]: <AxesSubplot:>
```



Graphique obtenu à partir du dataFrame filtre_indicateurs venant de sData

Le remplissage est-il satisfaisant pour tous les pays?

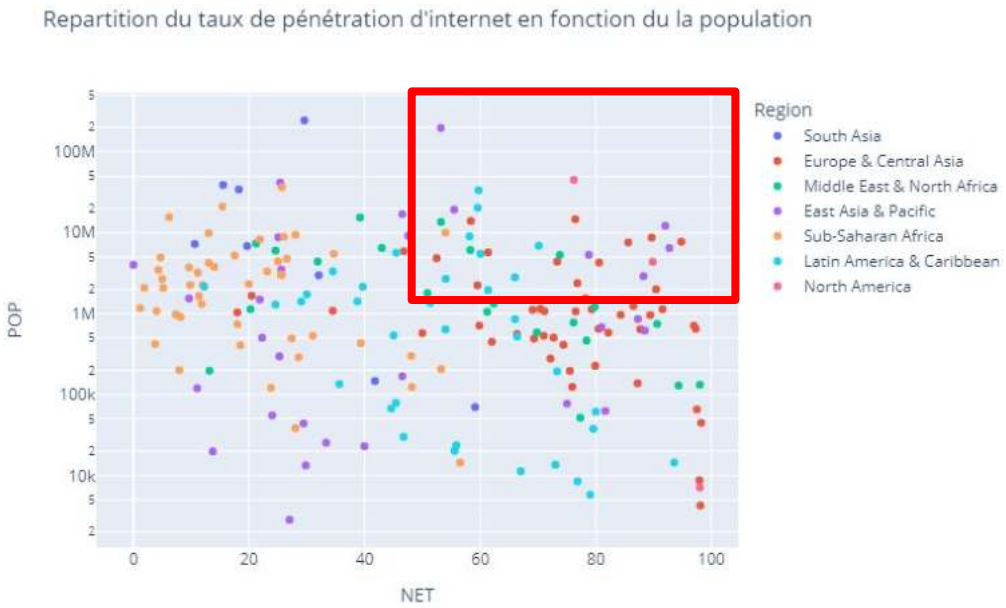


Graphique obtenu à partir du dataFrame data venant de sData

Pour 209 pays et groupes de pays nous avons un taux de remplissage à 95% environ.

Ces indicateurs sont exploitables. Nous pouvons commencer à chercher les pays prioritaires.

Aidons-nous d'une analyse croisée de deux indicateurs pour commencer à sélectionner des pays



Graphique obtenu à partir du dataFrame data_pivot venant de sData

Les pays qui vont nous intéresser pour l'expansion sont dans le cadre supérieur droit du graphique. cad NET >45% & POP > 2M

On peut notamment repérer :

- la Chine
- le Brésil
- Mexico
- USA
- la Russie
- le Japon

Un exemple de statistiques descriptives par régions pour le taux de pénétration d'internet

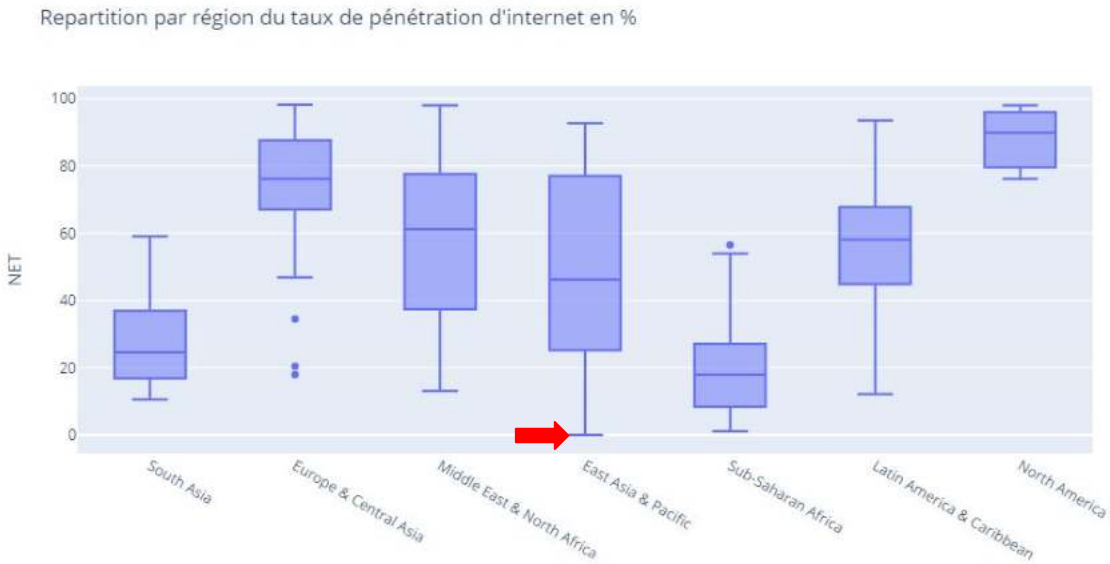
```
Net = data_pivot[['NET','Region']].groupby('Region',as_index = True).describe()
Net.reset_index(inplace = True)
Net
```

Region	NET							
	count	mean	std	min	25%	50%	75%	max
East Asia & Pacific	34.0	48.32344	28.26347	0.00000	25.27626	46.25458	76.50750	92.71655
Europe & Central Asia	54.0	74.31966	18.37495	17.99032	67.41785	76.15472	87.58707	98.24002
Latin America & Caribbean	37.0	55.99694	18.05718	12.23260	45.00000	58.13649	67.03000	93.54245
Middle East & North Africa	21.0	57.95744	25.71617	13.13492	39.21381	61.17838	77.28939	97.99998
North America	3.0	88.00558	11.02667	76.17674	83.00837	89.84000	93.92000	98.00000
South Asia	8.0	28.31367	16.02429	10.59573	17.56384	24.61796	34.48141	59.09259
Sub-Saharan Africa	47.0	20.31210	14.59060	1.17712	8.86802	18.00000	26.94361	56.51471

Il y a encore hélas de grandes disparités dans l'implantation d'internet pour certaines régions.

L'**Europe et l'Asie centrale** , l'**Amérique du Nord** et le **Moyent orient et l'Afrique du Nord** sont les régions où l'implantation d'internet est la plus élevée.

Les distributions sont variés pour l'implantation d'internet

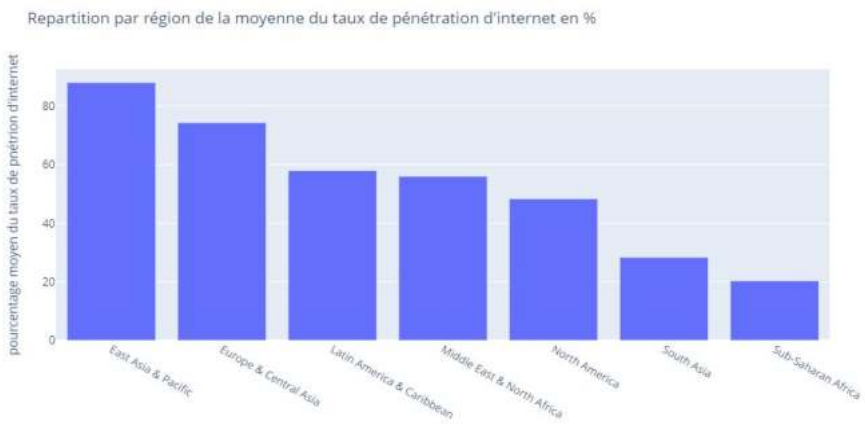


Graphique obtenu à partir du dataFrame data_pivot venant de sData

Pour l'Asie de l'Est et du Pacifique nous avons une grande dispersion des valeurs allant de 0 à 92%.

0 étant la Corée du Nord donc une valeur très probable.

Quels sont les régions ayant le plus de potentiels?

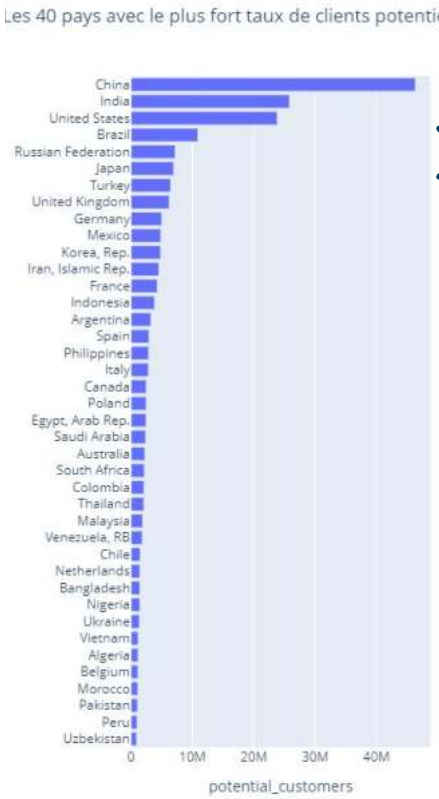


Voici un échantillon des graphiques obtenus pour deux de nos six indicateurs.
On remarque le même trio de tête pour chaque indicateur.

Top régions :

- Asie de l'Est et Pacifique
- Europe et Asie Centrale
- Amérique latine et Caribéenne

Application d'un filtre optimisé pour déterminer les pays à fort potentiels



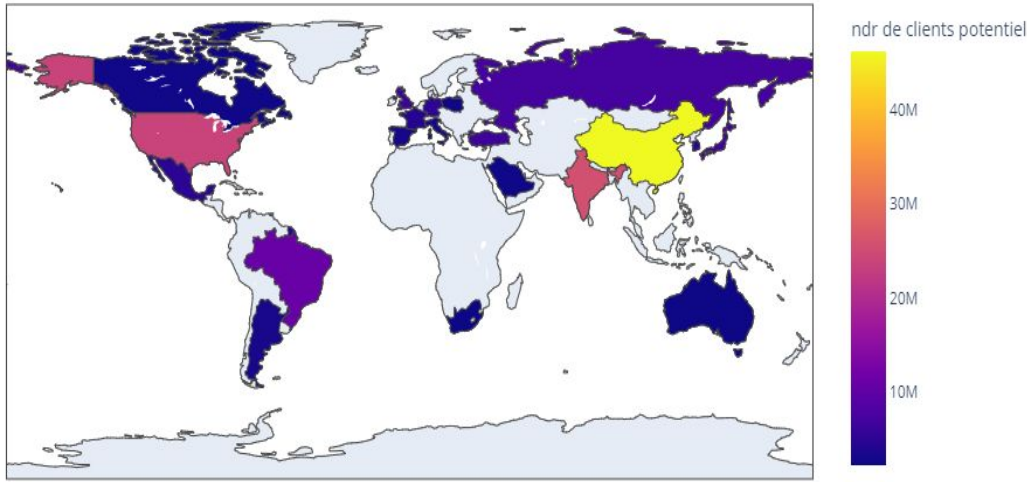
Pas assez de critères

NET >52%

Nb Clients potentiels >= 2,1M

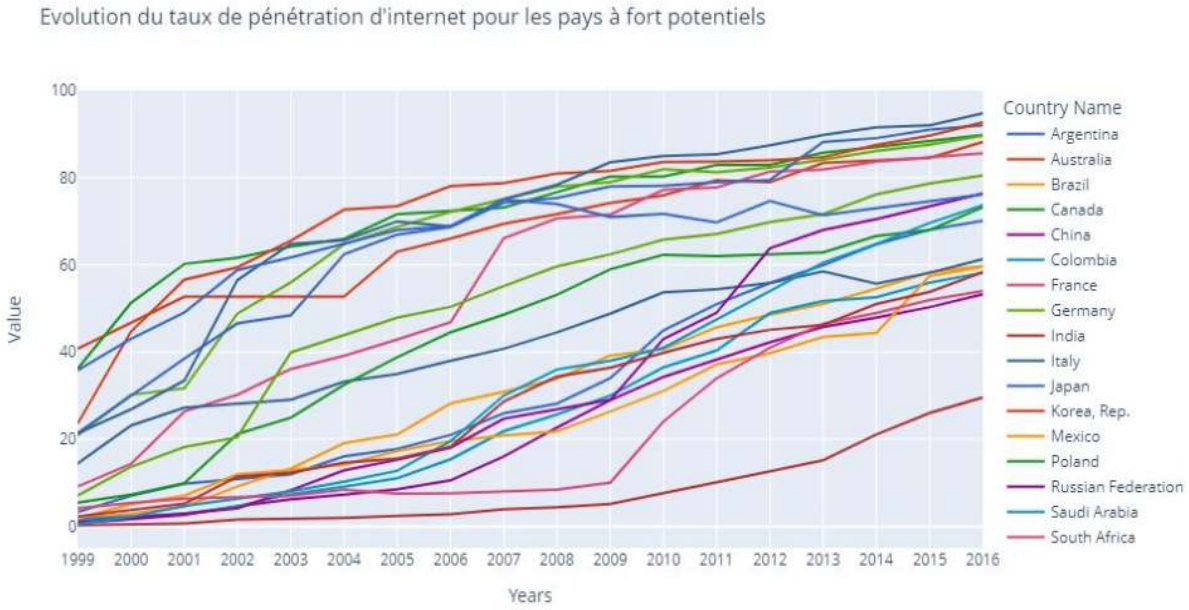
%PIB > 3%

20 pays prometteurs selon le nombre de clients potentiels



1. Chine
2. Inde
3. USA
4. Brésil
5. Russie
6. Japon
7. Turquie
8. Angleterre
9. Allemagne
10. Mexique
11. Corée
12. France
13. Argentine
14. Espagne
15. Italie
16. Canada
17. Pologne
18. Arabie Saoudite
19. Australie
20. Afrique du Sud

Pour chacun de ces pays, quelle sera l'évolution de l'accès à internet ?

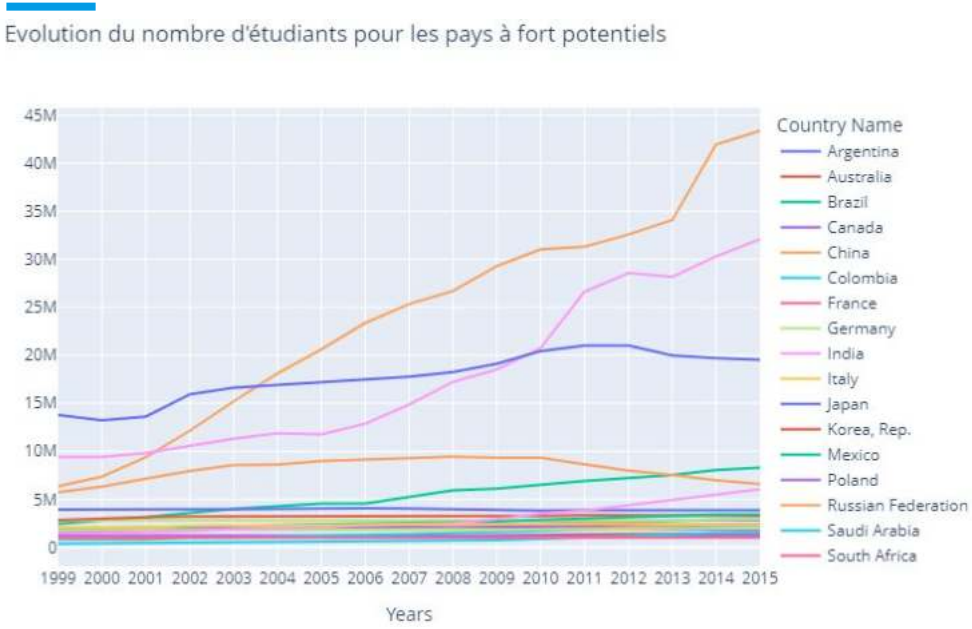


Graphique obtenu à l'aide de EdStatData

Pour l'implantation d'internet dans chaque pays

Nous avons une tendance croissante, continue et rapide pour quasiment tous les pays.

Pour chacun de ces pays, quelle sera l'évolution du nombre d'étudiants ?



Graphique obtenu à l'aide de EdStatData

On observe une croissance rapide et continue pour la Chine et l'Inde.

Nous avons une tendance à croissance lente voir stagnante pour les autres pays.

Globalement le nombre de jeunes de 15 à 24 ans stagne. De même pour le nombre d' élèves au lycée

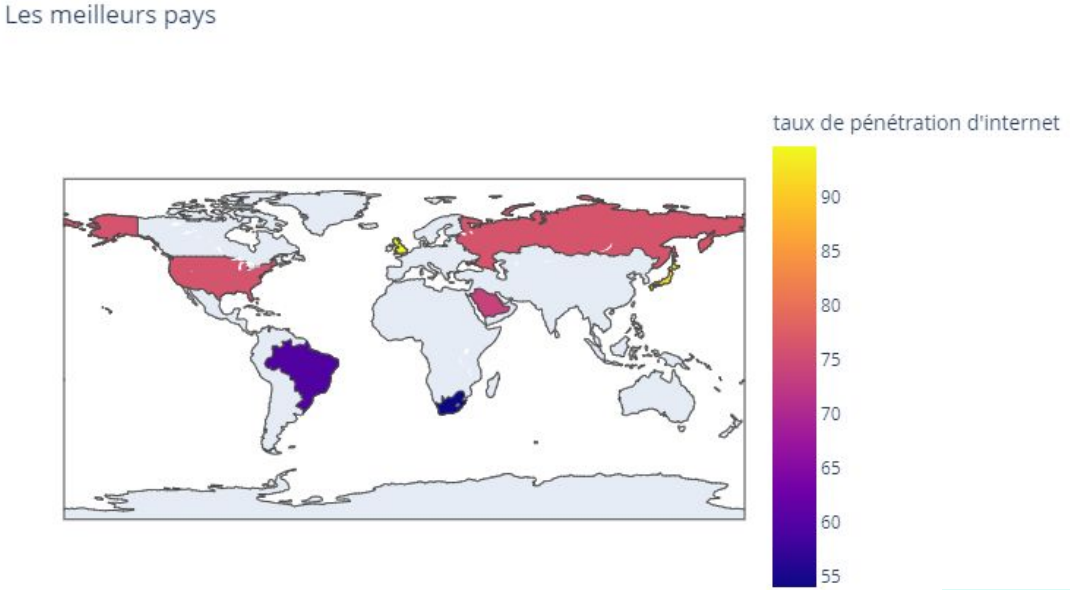
Conclusions

- Dans quels pays devons-nous opérer en priorité?
- Quels sont les pistes d’amélioration de cette étude?
- Question-réponses

Conclusion : Dans quels pays l'entreprise doit-elle opérer en priorité ?

1. Angleterre
2. USA
3. Russie
4. Brésil
5. Japon
6. Arabie Saoudite
7. Afrique du Sud

Choix personnel : Implantation en se positionnant dans chaque région du monde.



Conclusion

Le jeu de donnée bien que très dense nous permet bien de sélectionner les pays pour notre extension.

Mais reste à voir :

- L'étude de l'éventuelle concurrence surtout dans les pays développés
- La politique de l'entreprise envers les langues locales qui entraînerait une traduction du site et des supports de cours. Donc du temps pour tout mettre au point.
- La politique de l'entreprise envers les îles et petits pays (pop < 1M hab) car bien que la population soit faible l'accès à internet et l'investissement du gouvernement sont très encourageant.

Questions - Réponses

Merci de votre attention