Parcours OpenClassRooms

Data Scientist

P2 Données de systèmes éducatifs

Benoît DELORME
Création : 28/09/2020
Mise à jour : 12/10/2020

Sommaire

1) Problématique et présentation du jeu de données

- a) Enjeux
- b) Problématique
- c) Présentation du jeu de données

2) Pré-analyse

- a) Valider la qualité du jeu de données
- b) Sélection des indicateurs
- c) Ordre de grandeur des indicateurs statistiques

3) Conclusions

- a) Réponse aux enjeux
- b) Réponse à la problématique

4) Questions

1) Problématique et présentation du jeu de données

a) Enjeux

Une start-up de la EdTech propose des formations en ligne pour un public de niveau lycée et université.

Cette start-up souhaite s'exporter à l'international :

Quels sont les pays avec un fort potentiel de clients pour nos services ?

Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?

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

1) Problématique et présentation du jeu de données

b) Problématique

La start-up a repéré un jeu de données qui pourrait l’intéresser.

La mission est de déterminer si ce jeu de données peut aider le projet d’expansion.

Déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays.

1) Problématique et présentation du jeu de données

c) Présentation du jeu de données

Le fichier **EdStatsData** est le fichier le plus volumineux. Il contient, pour chaque pays, les valeurs des indicateurs par année.
886930 lignes et 70 colonnes.

Le fichier **EdStatsCountry** contient des informations générales sur les **pays**, leur niveau de richesse, les dates des dernières enquêtes, les catégories financières et politiques, … Il ne contient pas d'indicateur.
241 lignes et 32 colonnes.

Le fichier **EdStatsSeries** est le tableau des **indicateurs**, avec un certain nombre de colonnes : définition, unité de mesure, périodicité, méthode d'agrégation.
3665 lignes et 21 colonnes.

Le fichier **EdStatsSeries** renseigne l'origine des indicateurs nationaux (population du pays, taux de croissance, …)
613 lignes et 4 colonnes.

Dans le fichier **EdStatsFootnote** on retrouve l'incertitude liée aux indicateurs, selon le pays, et selon l'année.
643638 lignes et 5 colonnes.

2) Analyse pré-exploratoire

a) Valider la qualité du jeu de données

Un rapide parcours des DataFrame m'a amené à classer les données qui nous intéressent selon les catégories ci-dessous :

1. Nos **clients** : leur âge, leur niveau d'étude, le taux de chômage, le taux de croissance, ...

2. Leur capacité à avoir **accès aux formations** : l'alphabétisation, l'accès à internet, ...

3. Leur **pouvoir d'achat** : pays pauvre, pays riche, ...

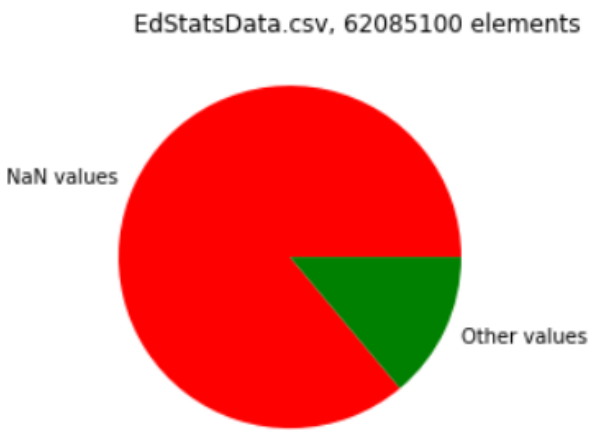
4. Les **conditions** dans lesquelles la start-up va opérer : le système économique, les éventuelles aides internationales, la stabilité politique du pays, ...

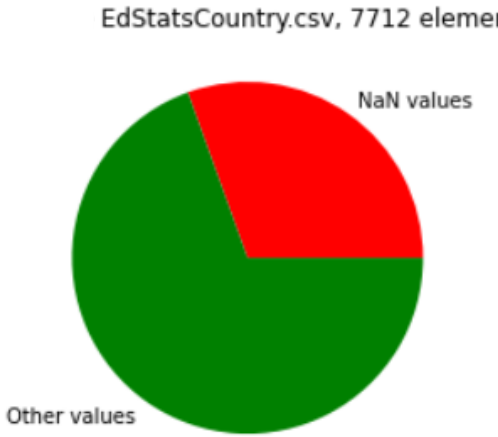
5. La **fiabilité** des données : l'ancienneté des enquêtes, recensement, ...

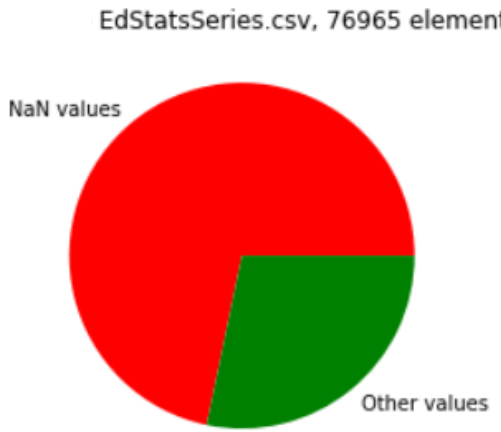
2) Analyse pré-exploratoire

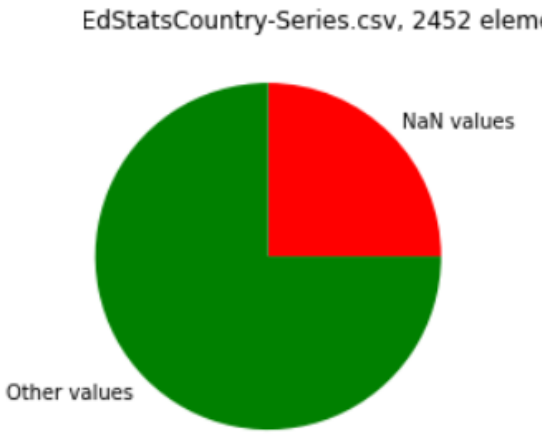
a) Valider la qualité du jeu de données

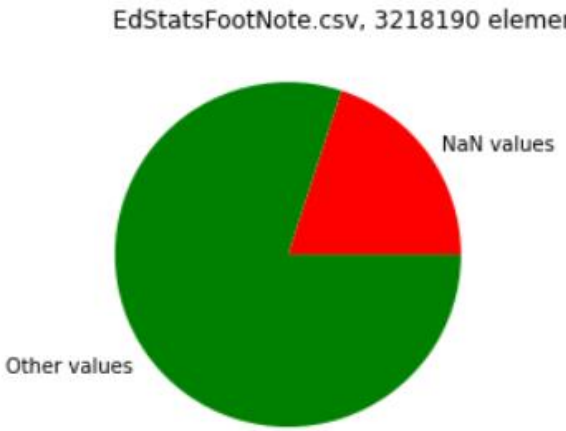
Données manquantes dans chaque fichier









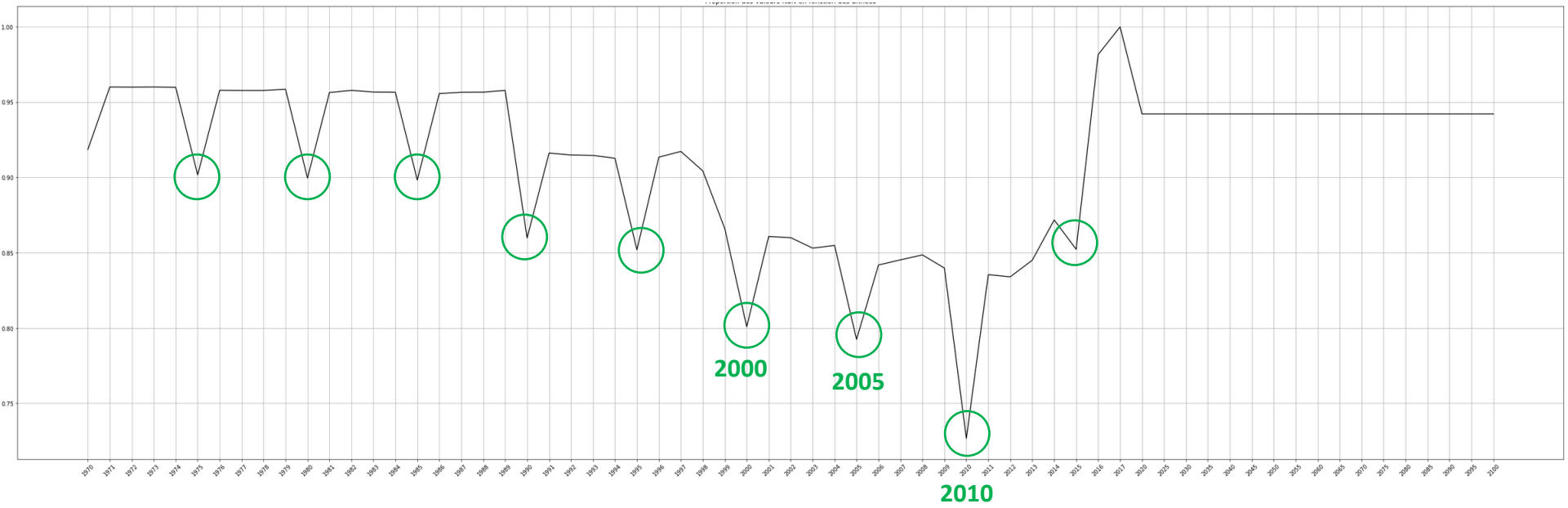


2) Analyse pré-exploratoire

a) Valider la qualité du jeu de données

Le fichier Data

Nombre de données non nulles par **années**



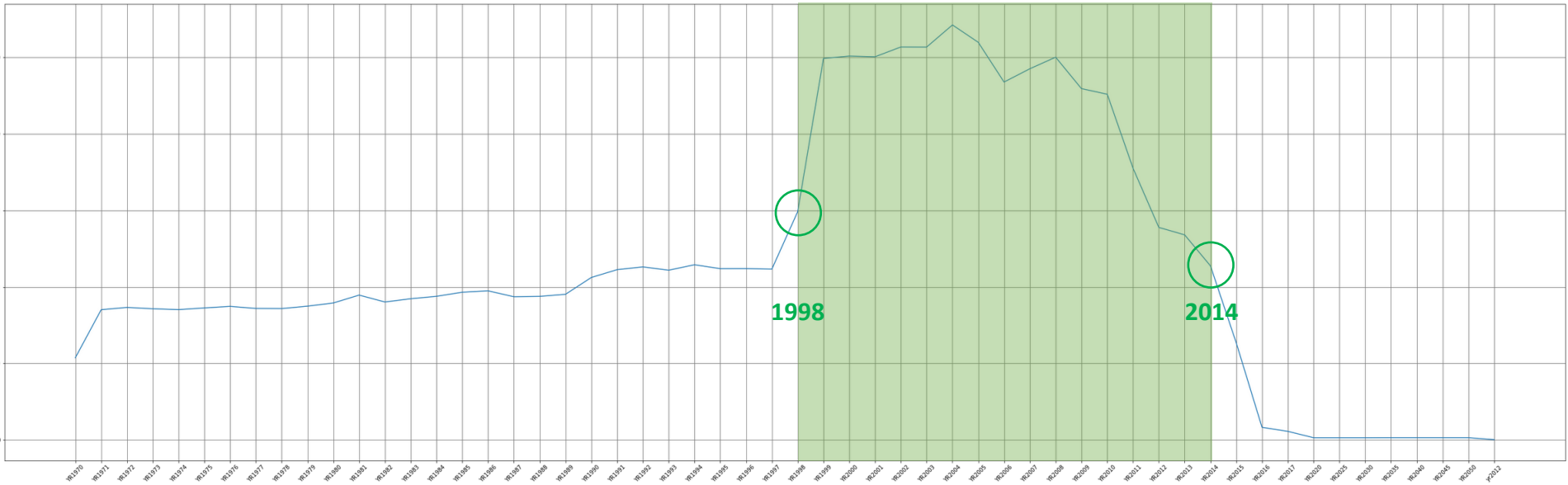
* La tendance générale montre une diminution des NaN de 1989 à 2015, puis une brusque augmentation.
* L'année la plus vide est 2017 (100% de valeurs NaN)
* L'année la plus remplie est 2010 (moins de 75% de NaN).
* Quelques pics sont bien visibles : de 1975 à 2010, tous les 5 ans, on remarque une diminution subite du nombre de NaN, au trement dit, un pic de données.
* On peut retenir 3 années pour lesquelles les NaN sont les moins nombreuses : 2000, 2005 et 2010.

2) Analyse pré-exploratoire

a) Valider la qualité du jeu de données

Le fichier Footnote

Nombre d'indicateurs disponibles par **année**



Dans l'analyse, il sera intéressant de se concentrer sur les années 1998 à 2014. Ce sont les années où la source de l'indicateur est connue.

2) Analyse pré-exploratoire

b) Sélection des indicateurs

Fichier Series : indicateurs

	Catégorie	Indicateur
0	Clients	Population, ages 10-18, total
1	Clients	Population, ages 15-24, total
2	Clients	Gross enrolment ratio, primary to tertiary, ge...
3	Clients	Enrolment in tertiary education per 100,000 in...
4	Clients	Unemployment, total (% of total labor force) (...
5	Clients	Population growth (annual %)
6	Accès aux formations	Youth literacy rate, population 15-24 years, b...
7	Accès aux formations	Personal computers (per 100 people)
8	Accès aux formations	Internet users (per 100 people)
9	Accès aux formations	Mortality rate, under-5 (per 1,000 live births)

Fichier Country

	Catégorie	Indicateur
0	Pouvoir d'achat	Income Group
1	Conditions d'activité	System of trade
2	Conditions d'activité	Lending category
3	Conditions d'activité	System of National Accounts
4	Conditions d'activité	Source of most recent Income and expenditure data

Fichier Series : évaluateurs

	Catégorie	Indicateur
0	Fiabilité des données	Periodicity
1	Fiabilité des données	Aggregation method
2	Fiabilité des données	Development relevance
3	Fiabilité des données	License Type

Fichier Country-Series

	Catégorie	Indicateur
0	Fiabilité des données	DESCRIPTION

2) Analyse pré-exploratoire

b) Sélection des indicateurs

Evaluation de la proportion de NaN dans les indicateurs retenus

	Catégorie	Indicateur	Nan Proportion
6	Accès aux formations	Youth literacy rate, population 15-24 years, b...	0.866234
7	Accès aux formations	Personal computers (per 100 people)	0.777922
2	Clients	Gross enrolment ratio, primary to tertiary, ge...	0.710094
0	Clients	Population, ages 10-18, total	0.660213
1	Clients	Population, ages 15-24, total	0.660213
3	Clients	Enrolment in tertiary education per 100,000 in...	0.639256
8	Accès aux formations	Internet users (per 100 people)	0.621783
5	Clients	Population growth (annual %)	0.284711

2) Analyse pré-exploratoire

c) Ordre de grandeur des indicateurs statistiques – années 2000, 2005, 2010

Population, ages 10-18

Données indisponibles !

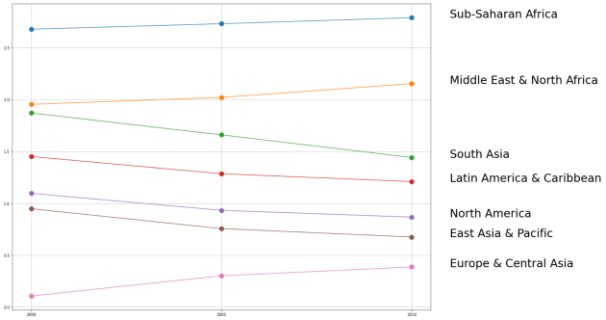
Enrolment in tertiary education

Données indisponibles !

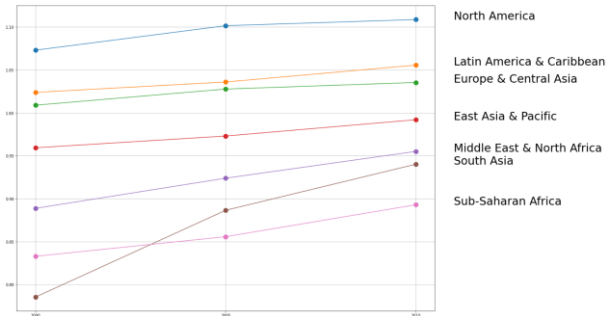
Population, ages 15-24

Données indisponibles !

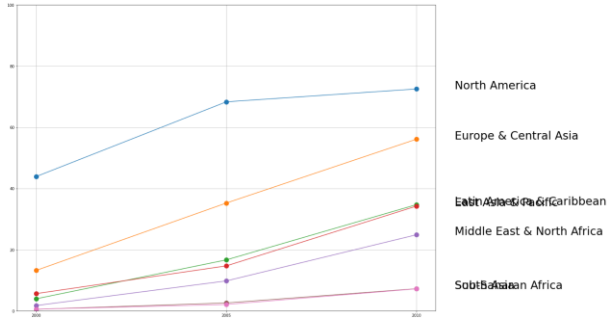
Population growth



Gross enrolment ratio, primary to tertiary



Internet users

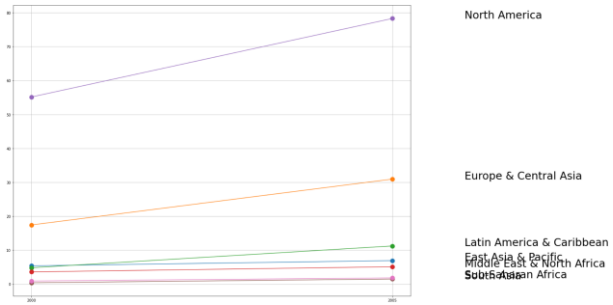


2) Analyse pré-exploratoire

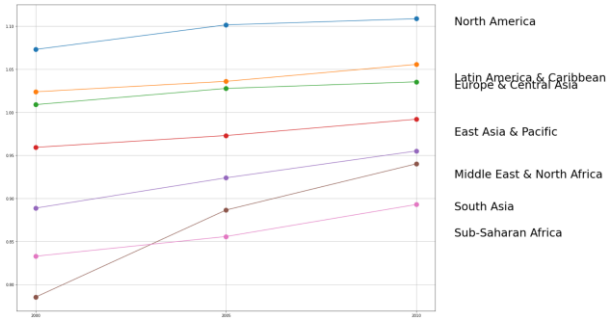
c) Ordre de grandeur des indicateurs statistiques – années 2020, 2025, 2030

Certains indicateurs écartés contiennent en fait des données exploitables

Personal computers (per 100 people)



Gross enrolment ratio, primary to tertiary, gender parity index (GPI)



P2_03_Conclusion : pertinence de l'usage du jeu de données

a) Réponse aux enjeux

Un jeu d'indicateur a pu être retenu, incluant divers catégories (nos clients, leur accès aux formations, …)

Certaines années sont plus fournies en données.

Les régions du monde présentant les meilleurs indicateurs sont les suivantes :
- Amérique du Nord
- Europe
- Latin America

b) Réponse à la problématique

Le jeu de données permet de tirer quelques conclusions sur les régions du monde les plus intéressantes.

Cela dit, plusieurs limites au modèle proposé :
- Les indicateurs ont été choisis en explorant manuellement le jeu de données, sans connaissance profonde de ce qu'ils représentent
- Parmi les indicateurs retenus, certains ne contiennent pas de données pour les années retenues, notamment les indicateurs de population. Or le continent asiatique est sujet à une forte croissance démographique, ce qui ne peut pas être pris en compte par le modèle actuel.
- Enfin, pour compléter l'étude, il serait intéressant d'exploiter les données de projection (2020 et après), concernant les populations et leurs éducation.

P2_04_Questions-réponses

Merci de votre attention !