Analyse exploratoire de données

Academy - Projet d’expansion

ACADEMY

10/10/2023

Introduction

- Academy propose des cours en ligne aux élèves de niveau lycée et université

- Contexte: projet d'expansion à l'international

- But de cette analyse exploratoire:
  - Déterminer si les données fournies permettent d'informer le projet d'expansion
  - Identifier les pays à fort potentiel de clients

Description du jeu de données

- Source: la Banque mondiale
- Plus de 4000 indicateurs relatifs à l'éducation (accès à l'éducation, l'alphabétisation, les enseignants, la population …) par pays
- Les indicateurs couvrent le cycle de l'éducation de la maternelle à l'enseignement supérieur.
- Ces données sont regroupées dans 5 fichiers:

Fichier	Nb de lignes	Nb de colonnes
EdStatsCountry	241	32
EdStatsCountrySeries	613	4
EdStatsData	886930	70
EdStatsFootNote	643638	5
EdStatsSeries	3665	21

Environnement Python

- Notebook Jupyter
- Création et activation d’un environnement virtuel “Projet2”

```
C:\Users\Vincent-Formation>python3 -m venv Projet2

C:\Users\Vincent-Formation>Projet2\Scripts\activate.bat

(Projet2) C:\Users\Vincent-Formation>
```

- Installation du package Missingno avec pip
- Vérification des versions des packages installés

```
(Projet2) C:\Users\Vincent-Formation>pip list
Package          Version
---------------- ------------
contourpy        1.1.1
cycler           0.12.1
fonttools        4.43.1
kiwisolver       1.4.5
matplotlib       3.8.0
missingno        0.5.2
numpy            1.26.0
packaging        23.2
pandas           2.1.1
Pillow           10.0.1
pip              23.2.1
pyparsing        3.1.1
python-dateutil  2.8.2
pytz             2023.3.post1
scipy            1.11.3
seaborn          0.13.0
setuptools       65.5.0
six              1.16.0
tzdata           2023.3
```

- Importation des librairies Numpy, Pandas, Matplotlib

Partie 1:
Nettoyage du jeu de données

Processus de nettoyage du jeu de données

Variables choisies

Indicator Name	Indicator Code	Description	Justification
Internet users (per 100 people)	IT.NET.USER.P2	Nombre d'utilisateurs d'internet pour 100 habitants.	Les formations en lignes nécessitent une connexion internet
Population, ages 15-24, total	SP.POP.1524.TO.UN	Nombre d'habitants âgés de 15 à 24 ans	Les pays ciblés devront compter suffisamment de jeunes
Enrolment in secondary education, both sexes (number)	SE.SEC.ENRL	Nombre d'élèves inscrits dans le secondaire	Les formations s'adressent à un public de niveau lycée et université
Enrolment in tertiary education, all programmes, both sexes (number)	SE.TER.ENRL	Nombre d'élèves inscrits dans l'enseignement supérieur	Les formations s'adressent à un public de niveau lycée et université
GDP per capita (current US$)	NY.GDP.PCAP.CD	Produit Intérieur Brut par habitant	Formations payantes. Le développement commercial nécessite un PIB/hab suffisant

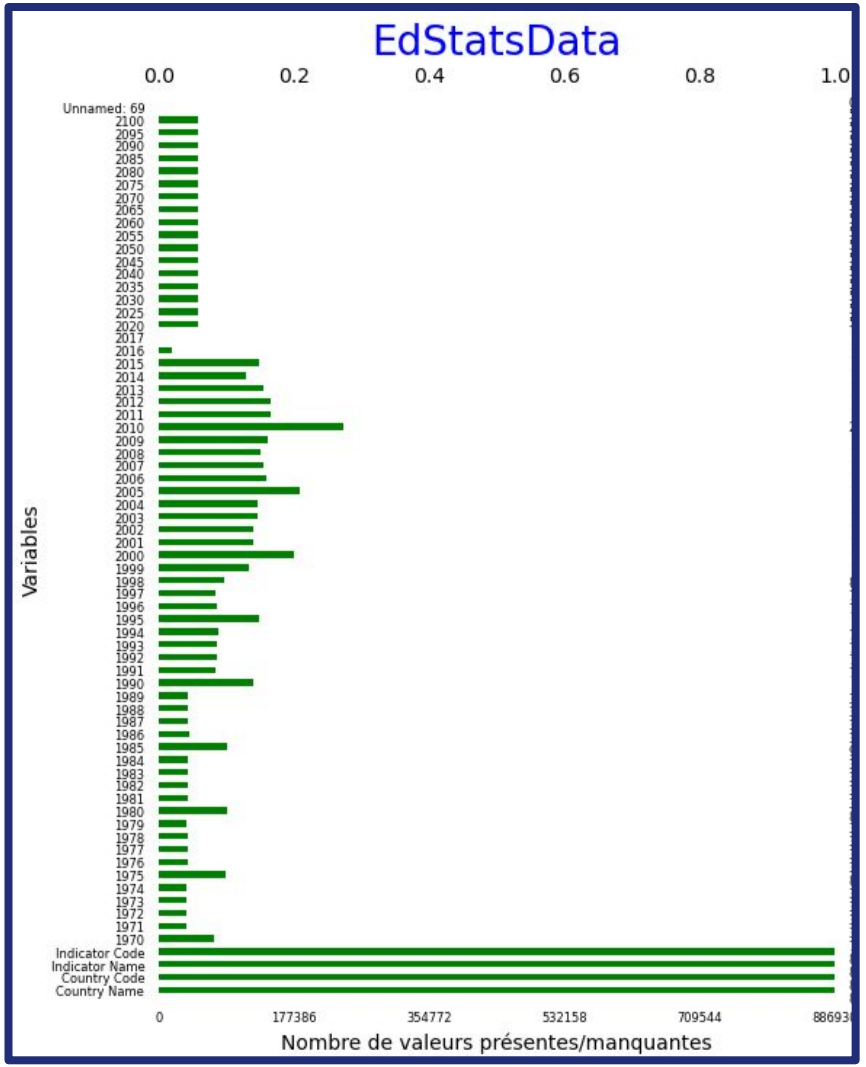
Qualité du jeu de données

- De nombreuses valeurs manquantes

  => visualisation Missingno

- Pas de doublons

- Plusieurs filtrages de données nécessaires



Traitement des erreurs

- Données très anciennes ou prospectives contiennent de nombreuses valeurs manquantes

- Traitement: script Python pour récupération de la dernière valeur non nulle des variables

Validation des données nettoyées

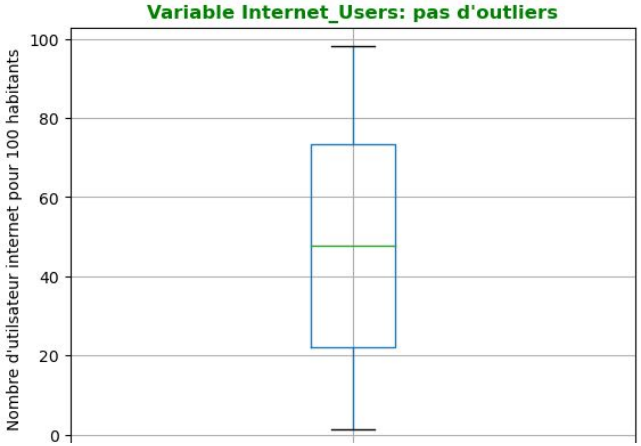
- Valeurs manquantes concernent majoritairement de tous petits pays qui ne sont pas pertinents

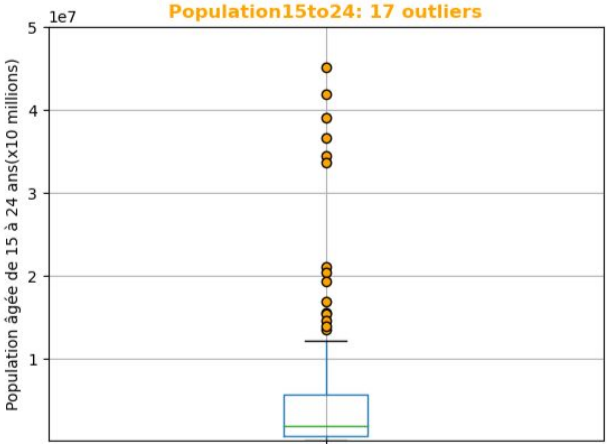
- Filtre sur population de 15-24 ans (130 000 hab minimum)

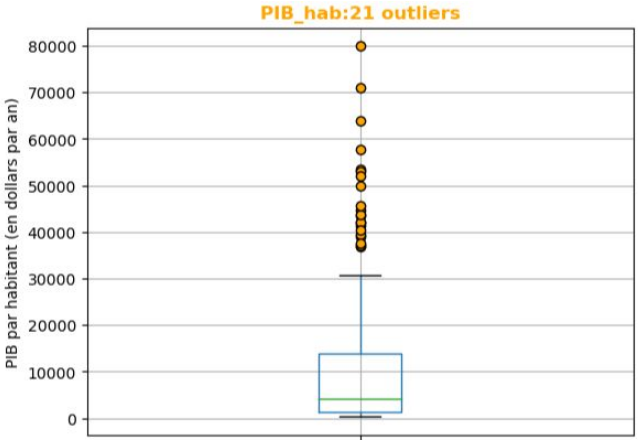
- En fixant ce seuil, les données sont exploitables pour les 5 variables choisies, pour 156 pays.

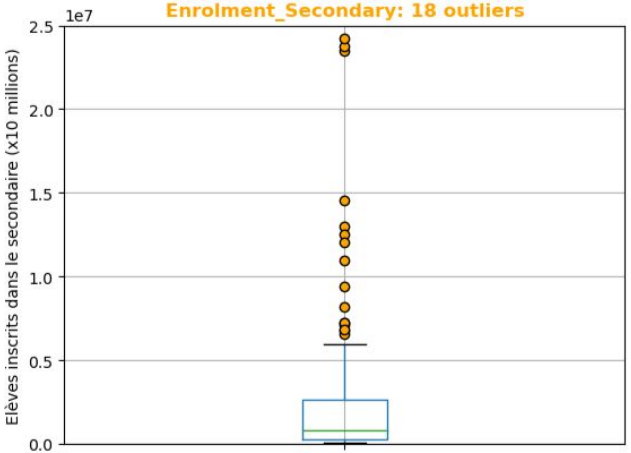
Partie 2:
Analyse exploratoire du jeu de données

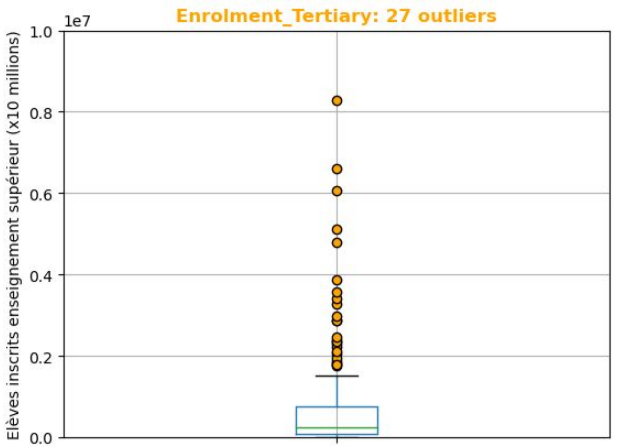
Détection d’outliers (approche statistique)











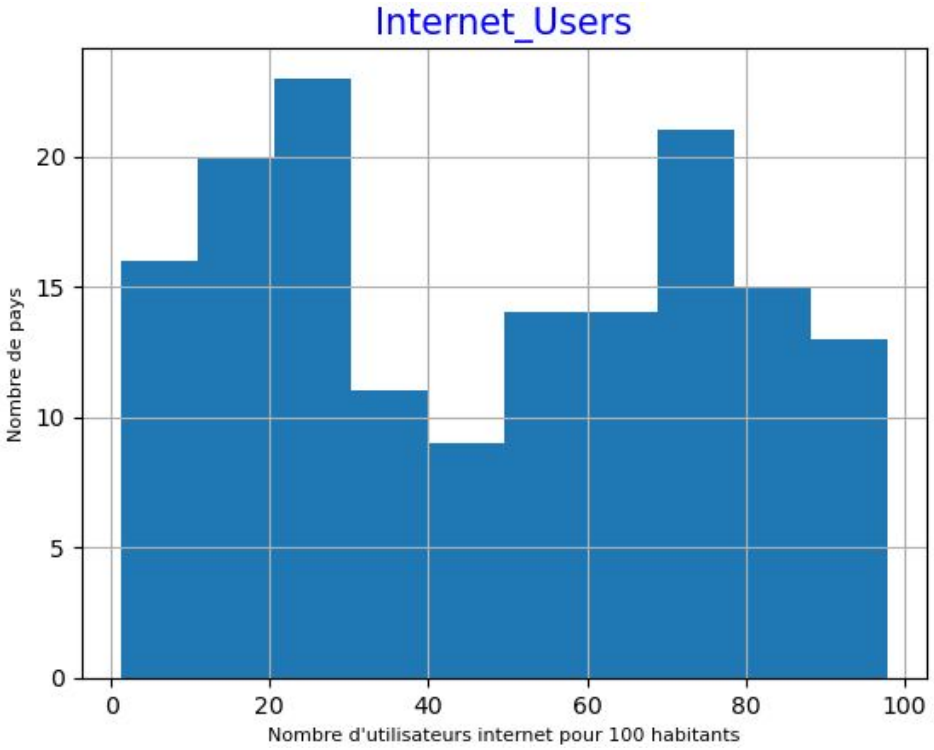
Détection d’outliers (approche métier)

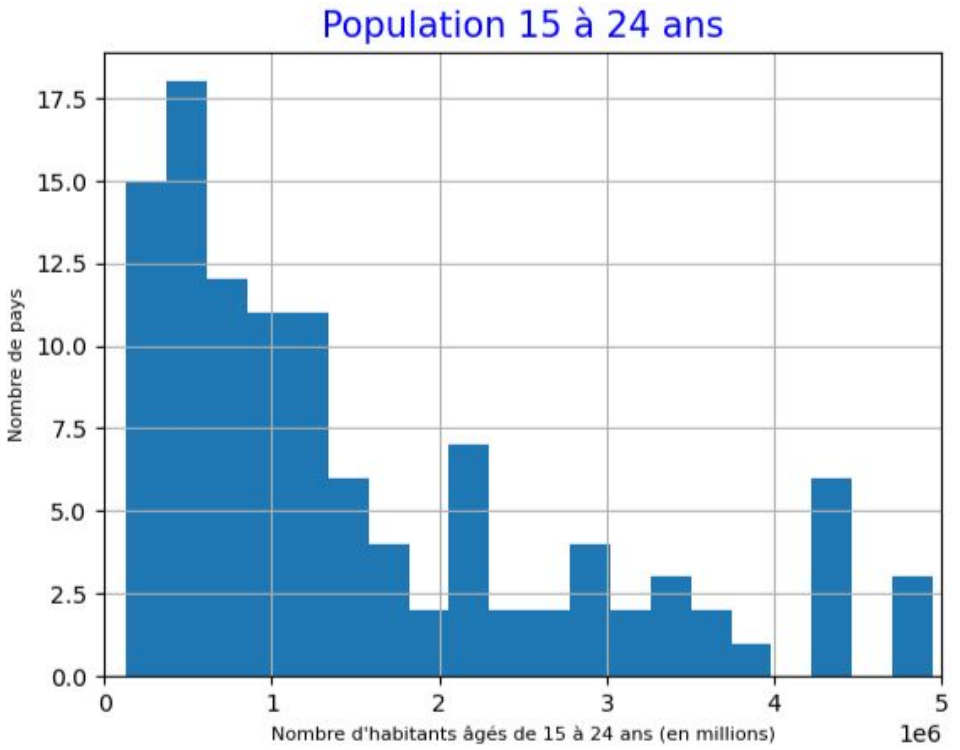
- Vérification du minimum et du maximum de chaque variable
- Nos outliers ne sont pas des valeurs aberrantes, mais des **valeurs atypiques**
- Ces valeurs atypiques sont réalistes et vérifiées, s’expliquent par une forte disparité par pays des populations et de la richesse
- Pas de traitement des outliers, **les valeurs atypiques doivent être conservées**

```
Data.describe()
```

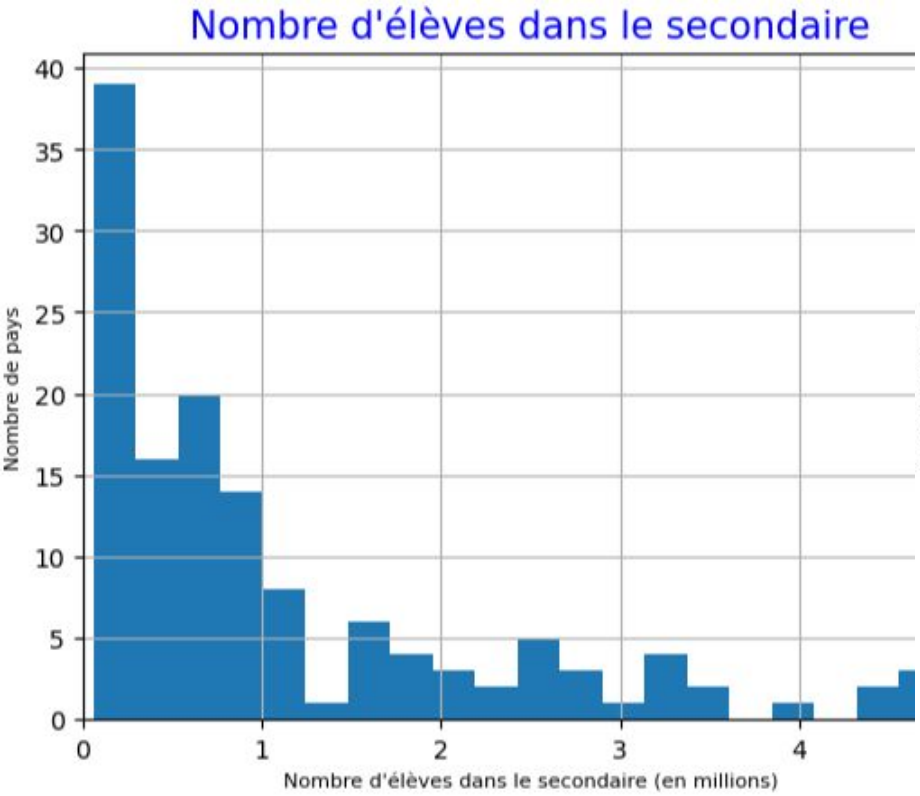
	Internet_Users	Population15to24	Enrolment_Secondary	Enrolment_Tertiary	PIB_hab
count	156.000000	1.560000e+02	1.560000e+02	1.560000e+02	156.000000
mean	47.596267	7.686036e+06	3.705644e+06	1.351753e+06	11612.918756
std	28.381641	2.575492e+07	1.281814e+07	4.655468e+06	16303.193158
min	1.177119	1.326090e+05	5.517600e+04	3.689000e+03	285.727442
25%	22.168370	7.022945e+05	3.068132e+05	9.370150e+04	1397.668952
50%	47.778620	1.984050e+06	8.021885e+05	2.583805e+05	4185.164354
75%	73.297881	5.666144e+06	2.614520e+06	7.611005e+05	13860.859728
max	97.999981	2.441202e+08	1.295421e+08	4.336739e+07	79890.524005

Analyse univariée (1/3)

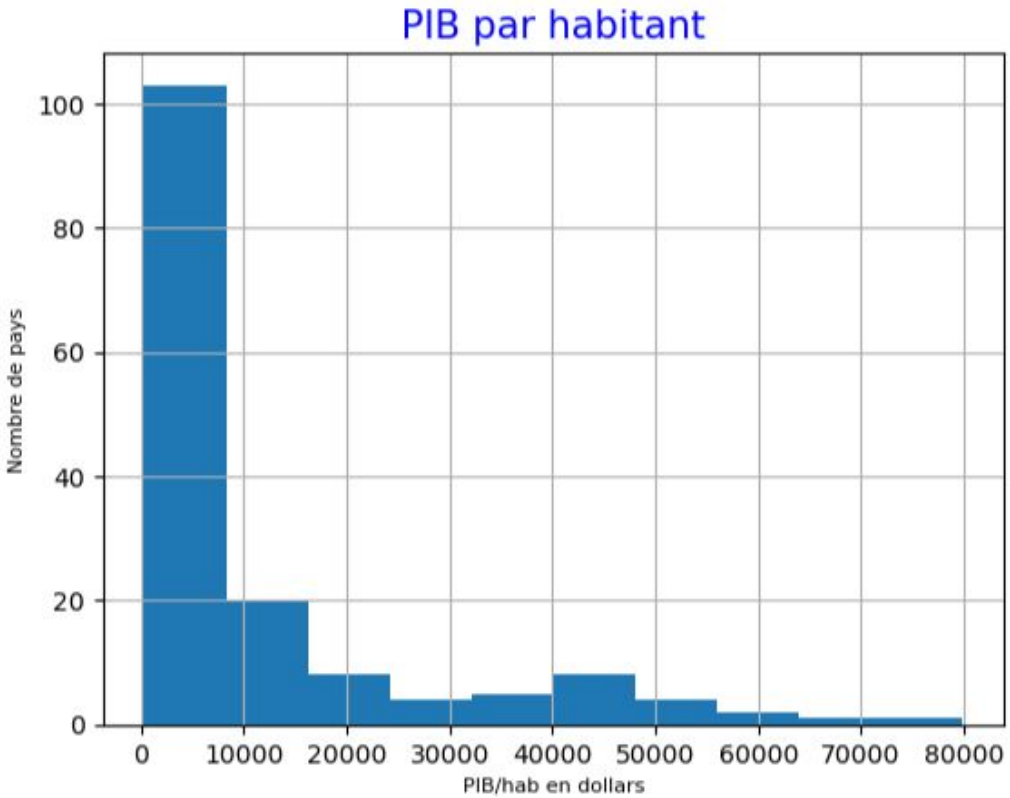




Analyse univariée (2/3)



Analyse univariée (3/3)



Score d’attractivité

- Score d’attractivité calculé à partir des 5 variables choisies
- Afin de comparer et d'analyser plus facilement les indicateurs, une normalisation est nécessaire.
- Choix d’une normalisation Z-score permet d'avoir une moyenne à 0 et un écart type à 1.
- Le calcul du score d’attractivité nous a permis d’établir un classement des pays à fort potentiel

Country Name	Attractivite
India	24.438633
China	22.863502
United States	10.825636
Switzerland	4.947523
Norway	4.642309
Brazil	4.297671
Japan	4.239565
Germany	4.016435
United Kingdom	3.876175
Ireland	3.653148
Denmark	3.583612
Australia	3.525960
Sweden	3.356146
France	3.291726
Canada	3.134790
Netherlands	3.114374
Singapore	2.944122
Korea, Rep.	2.915786
Russian Federation	2.687716
Austria	2.635523

Conclusion

- Les données, même si elles comportent de nombreuses valeurs manquantes permettent d'informer le projet d'expansion à l'international
- Le calcul du score d'attractivité nous a permis d'établir un classement des pays à fort potentiel