Parcours Data Scientist

SOUTENANCE PROJET 2 : ANALYSEZ DES DONNÉES DE SYSTÈMES ÉDUCATIFS

10 mars 2022

Présentation

Gauthier RAULT
Data Scientist chez academy
start-up de la EdTech

academy

Edwards DEMING:
“Without data you're just a person with an opinion.”
(Sans données, vous êtes juste quelqu'un avec une opinion)

Ordre du jour

- 5 min - Introduction, rappel de la problématique et présentation du jeu de données
- 15 min - Présentation de l'analyse pré-exploratoire du jeu de données et conclusions sur la pertinence de l'usage de ce dernier pour répondre aux questions stratégiques que se pose academy
- 5 à 10 min - Questions-réponses

1. Introduction (5min)

Contexte et présentation du jeu de données

Contexte

academy :

- ➔ propose du e-learning (formation en ligne)
- ➔ public de niveau lycée et université
- ➔ souhaite s'étendre à l'international

En s'appuyant sur les données Banque mondiale, nous répondrons aux questions:

- ★ Quels sont les pays avec un fort potentiel de clients pour leurs services ?
- ★ Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- ★ Dans quels pays doivent – ils s'implanter en priorité ?

Présentation du jeu de données

5 fichiers csv à disposition :

1. EdStatsCountry-Series.csv
2. EdStatsCountry.csv
3. EdStatsData.csv
4. EdStatsFootNote.csv
5. EdStatsSeries.csv

- Granularité nationale et annuelle
- de 1970 à 2100
- 1ère publication en 2010
- Dernière MAJ mars 2020
- MAJ trimestrielle (février, juin, août et novembre)

LA BANQUE MONDIALE

Source : *https://datacatalog.worldbank.org/dataset/education-statistics*

Phase d’inspection du jeu de données à disposition

- Quantitatif
- Qualitatif
- Forme (possibilité de croiser les documents)

→ Objectif

Se donner une première impression sur l'ensemble des documents en systématisant la démarche comme suit:

- *#print(x)*
- *#x.info()*
- *#x.head(2)*
- *#print((x.isnull().mean()).sort_values(ascending=[False]))*
- *#x.duplicated(keep=False).sum()*
- *#x.describe()*
- *#msno.matrix(x[x.columns[:]])*

EdStatsCountry-Series

- Ce fichier contient la source des données pour les informations contenues dans EdStatsCountry.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 613 entries, 0 to 612
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   CountryCode   613 non-null    object
 1   SeriesCode    613 non-null    object
 2   DESCRIPTION   613 non-null    object
 3   Unnamed: 3    0 non-null      float64
dtypes: float64(1), object(3)
memory usage: 19.3+ KB
```

- Quantité de données :
  - *613 lignes*
  - *4 colonnes*
- Qualité du fichier :
  - *Pas de doublon*
  - *Complet sauf colonne Unnamed: 3*

	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
1	ABW	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
2	AFG	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
3	AFG	NY.GDP.PCAP.PP.CD	Estimates are based on regression.	NaN
4	AFG	SP.POP.TOTL	Data sources : United Nations World Population...	NaN

EdStatsCountry

- Ce fichier contient des informations globales sur chaque pays du monde

- Quantité de données :
  - *241 lignes*
  - *32 colonnes*
- Qualité du fichier :
  - *Pas de doublon*
  - *Bonne mais valeur manquantes suivant les colonnes*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241 entries, 0 to 240
Data columns (total 32 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   Country Code                                          241 non-null    object
 1   Short Name                                            241 non-null    object
 2   Table Name                                            241 non-null    object
 3   Long Name                                             241 non-null    object
 4   2-alpha code                                          238 non-null    object
 5   Currency Unit                                         215 non-null    object
 6   Special Notes                                         145 non-null    object
 7   Region                                                214 non-null    object
 8   Income Group                                          214 non-null    object
 9   WB-2 code                                             240 non-null    object
 10  National accounts base year                           205 non-null    object
 11  National accounts reference year                      32 non-null     float64
 12  SNA price valuation                                   197 non-null    object
 13  Lending category                                      144 non-null    object
 14  Other groups                                          58 non-null     object
 15  System of National Accounts                           215 non-null    object
 16  Alternative conversion factor                         47 non-null     object
 17  PPP survey year                                       145 non-null    object
 18  Balance of Payments Manual in use                     181 non-null    object
 19  External debt Reporting status                        124 non-null    object
 20  System of trade                                       200 non-null    object
 21  Government Accounting concept                         161 non-null    object
 22  IMF data dissemination standard                       181 non-null    object
 23  Latest population census                              213 non-null    object
 24  Latest household survey                               141 non-null    object
 25  Source of most recent Income and expenditure data     160 non-null    object
 26  Vital registration complete                           111 non-null    object
 27  Latest agricultural census                            142 non-null    object
 28  Latest industrial data                                107 non-null    float64
 29  Latest trade data                                     185 non-null    float64
 30  Latest water withdrawal data                          179 non-null    object
 31  Unnamed: 31                                           0 non-null      float64
dtypes: float64(4), object(28)
memory usage: 60.4+ KB
```

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	IMF data dissemination standard	Latest population census	Latest household survey	Source o most recen Income an expenditur dat
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD	AW	...	NaN	2010	NaN	Nat
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income	AF	...	General Data Dissemination System (GDDS)	1979	Multiple Indicator Cluster Survey (MICS), 2010/11	Integrate household survey (IHS, 200

EdStatsFootNote

- Ce fichier contient des informations sur l'année d'origine et les incertitudes des données

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 643638 entries, 0 to 643637
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype  
---  ------        --------------   -----  
 0   CountryCode   643638 non-null  object 
 1   SeriesCode    643638 non-null  object 
 2   Year          643638 non-null  object 
 3   DESCRIPTION   643638 non-null  object 
 4   Unnamed: 4    0 non-null       float64
dtypes: float64(1), object(4)
memory usage: 24.6+ MB
```

- Quantité de données :
  - *643 638 lignes*
  - *5 colonnes*
- Qualité du fichier :
  - *Pas de doublon*
  - *Complet sauf colonne Unnamed: 4*

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN

EdStatsSeries

- Ce fichier contient des informations sur les indicateurs du fichier EdStatsData - doc3

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3665 entries, 0 to 3664
Data columns (total 21 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Series Code                           3665 non-null   object
 1   Topic                                 3665 non-null   object
 2   Indicator Name                        3665 non-null   object
 3   Short definition                      2156 non-null   object
 4   Long definition                       3665 non-null   object
 5   Unit of measure                       0 non-null      float64
 6   Periodicity                           99 non-null     object
 7   Base Period                           314 non-null    object
 8   Other notes                           552 non-null    object
 9   Aggregation method                    47 non-null     object
 10  Limitations and exceptions            14 non-null     object
 11  Notes from original source            0 non-null      float64
 12  General comments                      14 non-null     object
 13  Source                                3665 non-null   object
 14  Statistical concept and methodology   23 non-null     object
 15  Development relevance                 3 non-null      object
 16  Related source links                  215 non-null    object
 17  Other web links                       0 non-null      float64
 18  Related indicators                    0 non-null      float64
 19  License Type                          0 non-null      float64
 20  Unnamed: 20                           0 non-null      float64
dtypes: float64(6), object(15)
memory usage: 601.4+ KB
```

- Quantité de données :
  - *3665 lignes*
  - *21 colonnes*
- Qualité du fichier :
  - *Pas de doublon*
  - *Peu de valeur renseignées*

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	Base Period	Other notes	Aggregation method	...	Notes from original source	General comments
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 15-19 with...	Percentage of female population age 15-19 with...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
1	BAR.NOED.1519.ZS	Attainment	Barro-Lee: Percentage of population age 15-19 ...	Percentage of population age 15-19 with no edu...	Percentage of population age 15-19 with no edu...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN

EdStatsData

- Ce fichier contient l'évolution par années des indicateurs par pays

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 70 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Country Name    886930 non-null  object
 1   Country Code    886930 non-null  object
 2   Indicator Name  886930 non-null  object
 3   Indicator Code  886930 non-null  object
 4   1970            72288 non-null   float64
 5   1971            35537 non-null   float64
 6   1972            35619 non-null   float64
 7   1973            35545 non-null   float64
 8   1974            35730 non-null   float64
 9   1975            87306 non-null   float64
 10  1976            37483 non-null   float64
```

- Quantité de données :
  - *886 930 lignes*
  - *70 colonnes*
- Qualité du fichier :
  - *Pas de doublon*
  - *Beaucoup de valeurs manquantes*

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2060	2065	2070	2075	2080	2085	209
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Nal
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Nal

2. Analyse pré-exploratoire (15 min)

Constat de la phase d’inspection

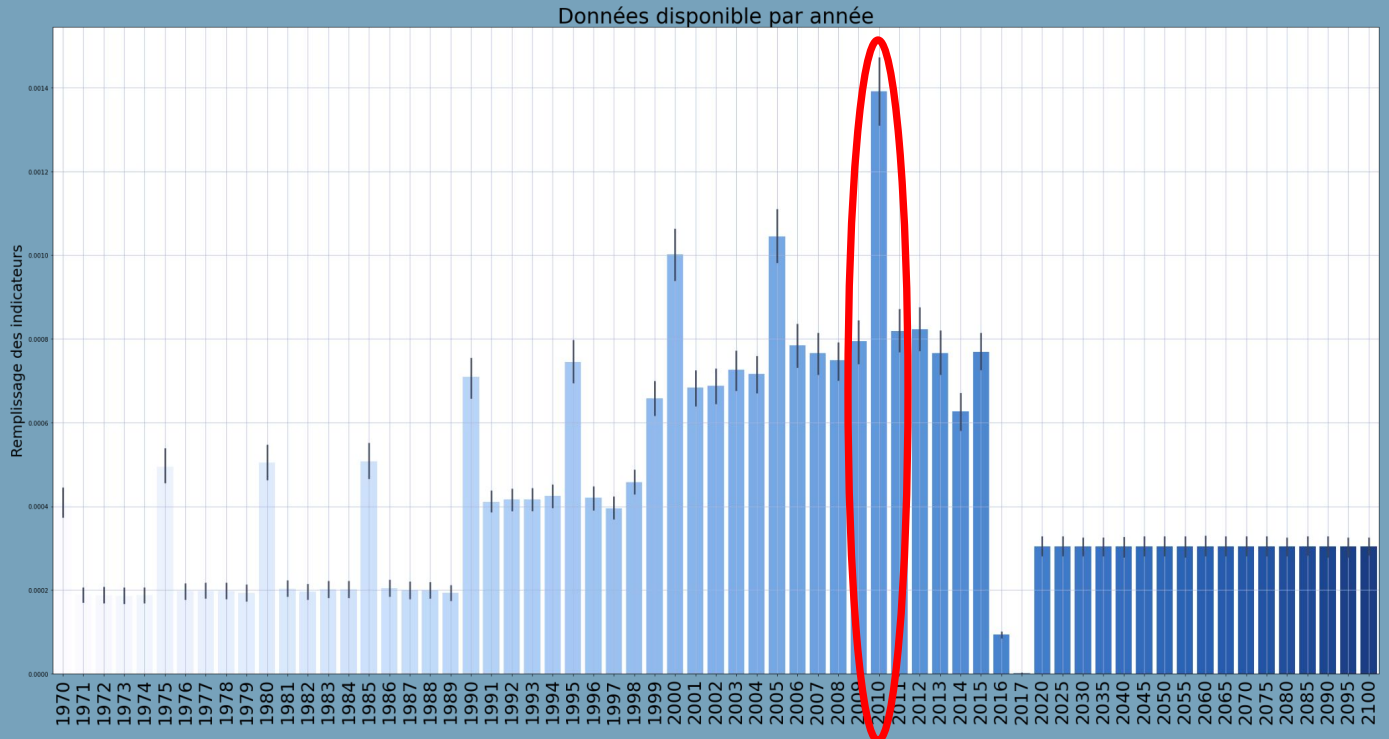
- 2 documents sont intéressants pour l’étude :
  - *EdStatsCountry "Country"*
  - *EdStatsData "Data"*

- Données inexploitables en l'état (grand nombre de données inconsistantes pour l’étude)

- Beaucoup de valeurs manquantes

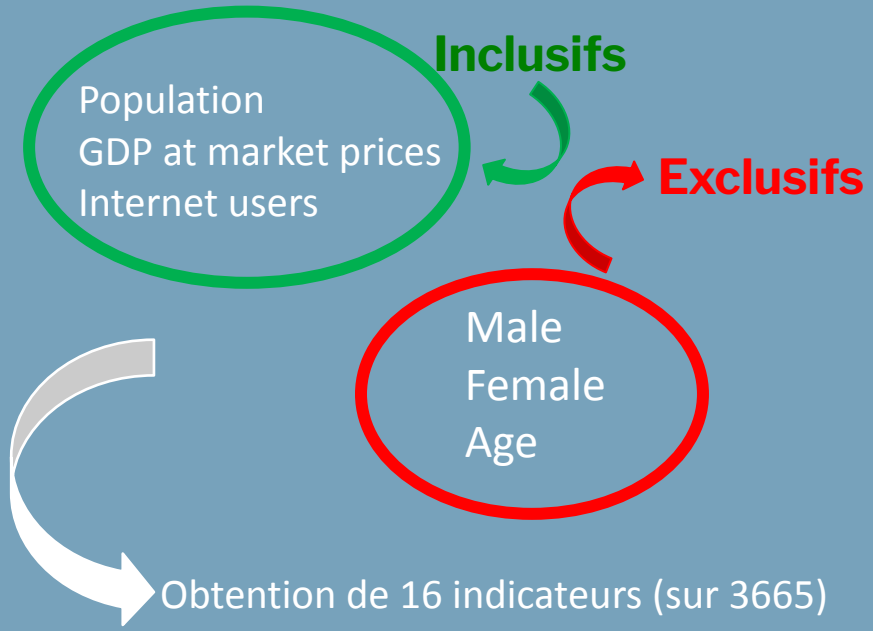
Stratégie de l'entonnoir pour se focaliser sur les données qui ont un intérêt pour l’étude.

Filtrage des années



Filtrage des indicateurs

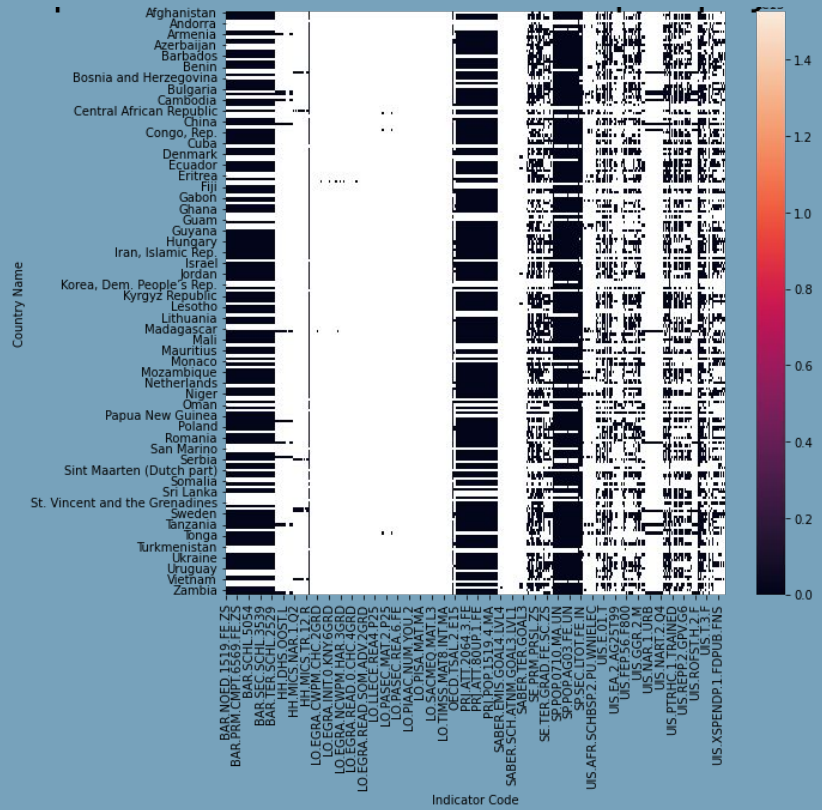
Mise en place d'une fonction basée sur une sélection de mots clefs :



Indicator Name	Country Name	Country Code	Indicator Code	2010
Population growth (annual %)	100.0	100.0	100.0	100.000000
Population, total	100.0	100.0	100.0	100.000000
GDP at market prices (current US$)	100.0	100.0	100.0	98.101266
Internet users (per 100 people)	100.0	100.0	100.0	98.101266
GDP at market prices (constant 2005 US$)	100.0	100.0	100.0	97.468354
Enrolment in secondary general, both sexes (number)	100.0	100.0	100.0	79.113924
Enrolment in secondary education, both sexes (number)	100.0	100.0	100.0	75.316456
Enrolment in tertiary education, all programmes, both sexes (number)	100.0	100.0	100.0	75.316456
Enrolment in tertiary education per 100,000 inhabitants, both sexes	100.0	100.0	100.0	73.417722
Enrolment in secondary education, public institutions, both sexes (number)	100.0	100.0	100.0	70.253165
Enrolment in secondary education, private institutions, both sexes (number)	100.0	100.0	100.0	66.455696
Enrolment in secondary vocational, both sexes (number)	100.0	100.0	100.0	66.455696
Enrolment in tertiary education, ISCED 5 programmes, both sexes (number)	100.0	100.0	100.0	57.594937
Enrolment in tertiary education, ISCED 8 programmes, both sexes (number)	100.0	100.0	100.0	53.164557
Enrolment in tertiary education, ISCED 6 programmes, both sexes (number)	100.0	100.0	100.0	1.898734
Enrolment in tertiary education, ISCED 7 programmes, both sexes (number)	100.0	100.0	100.0	1.265823

Synthèse des complétudes pour 2010

Heatmap - Complétude des indicateurs par pays en 2010



Sélection des indicateurs utiles

6 indicateurs sélectionnés associant:

- Cohérence pour l'étude
  - *Accès à internet*
  - *Richesse*
  - *Population*
  - *Nombre de lycéens et universitaires*

- Taux de complétion

```
'SP.POP.TOTL',
'SP.POP.GROW',
'NY.GDP.MKTP.CD',
'IT.NET.USER.P2',
'SE.SEC.ENRL.GC',
'SE.TER.ENRL',
```

Filtre sur les pays

2 filtres appliqués pour se focaliser sur les pays ayant de l'intérêt dans notre étude:

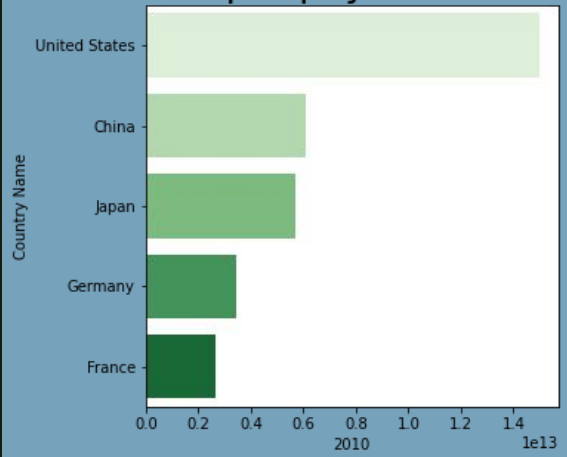
1. n'appartenant pas à une région (ex World)
2. < à 1 million d'habitants

=

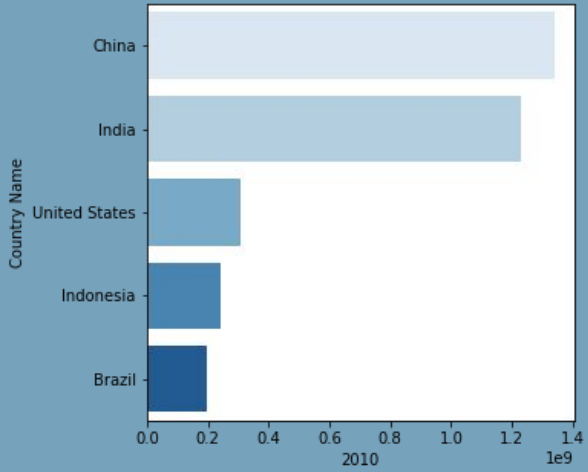
158 pays

Stratégie du top 5 par pays en 2010

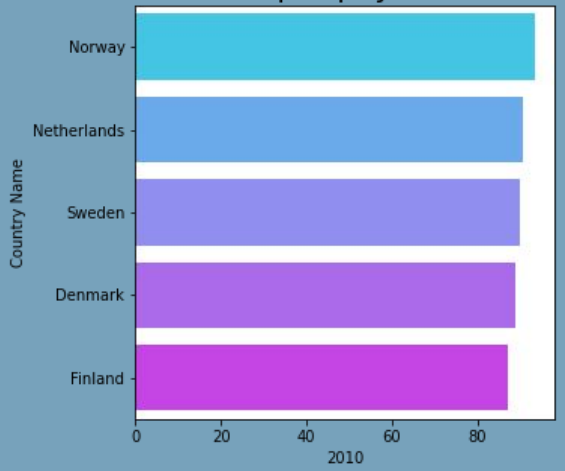
PIB



Population

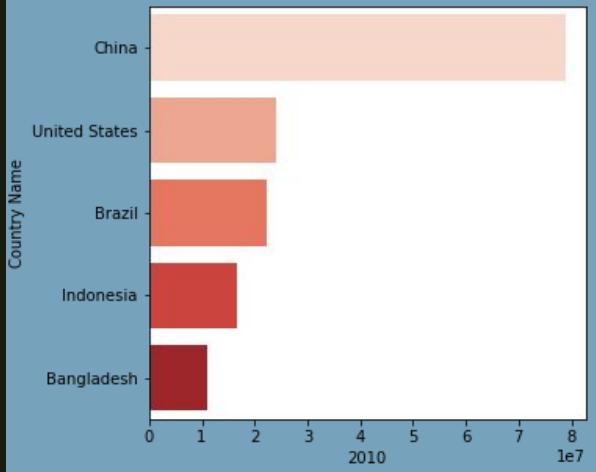


Internet

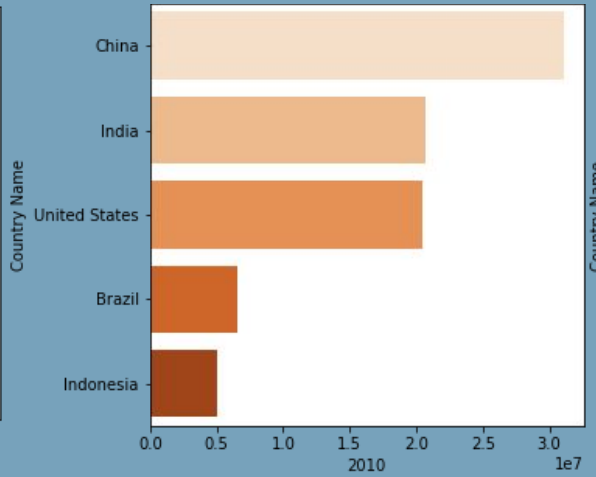


Stratégie du top 5 par pays en 2010

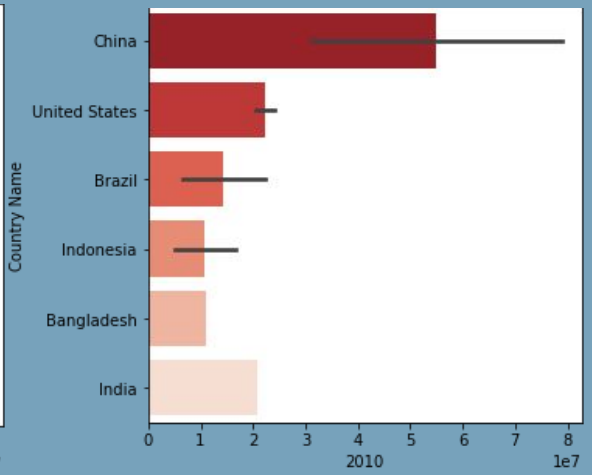
Secondaire



3ème cycle

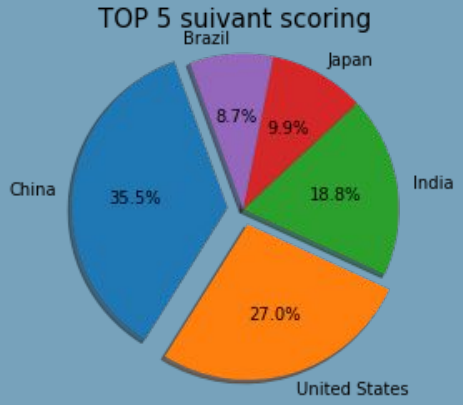


Supérieur



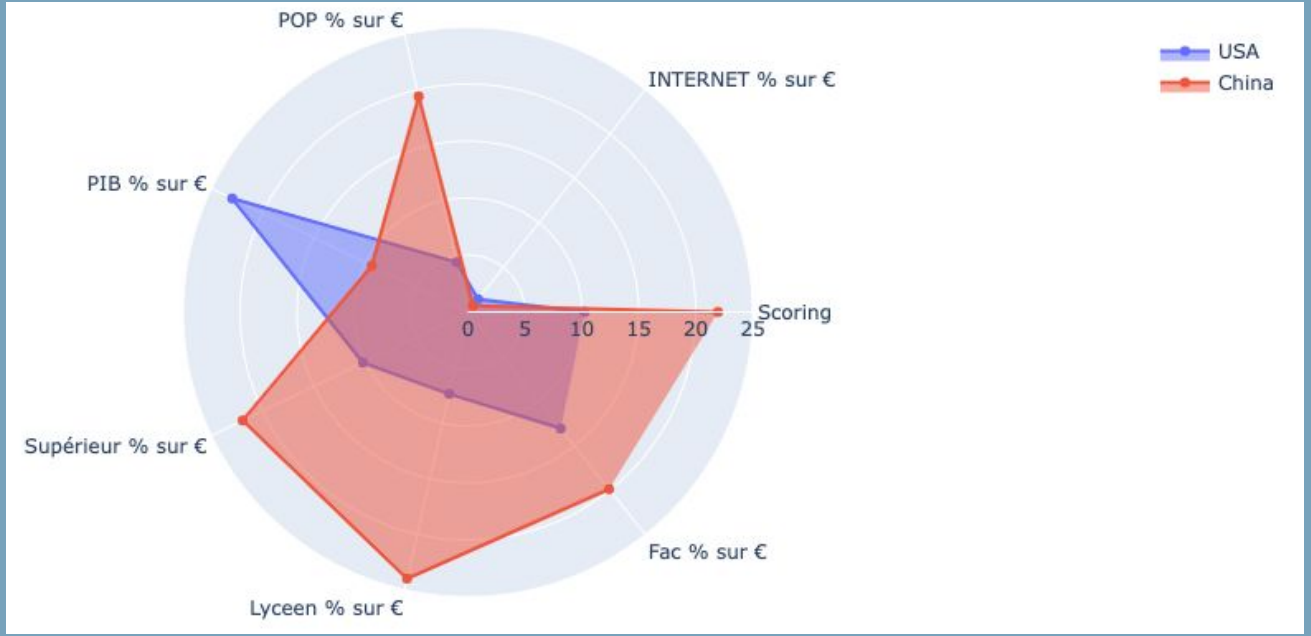
Scolarité par pays en 2010

Stratégie du top 5 par pays en 2010

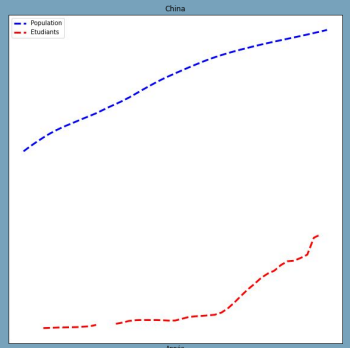


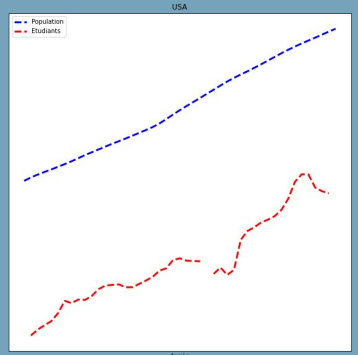
Pays	INTERNET % sur €	POP % sur €	PIB % sur €	Supérieur % sur €	Fac % sur €	Lyceen % sur €	Scoring
China	0.688845	19.406024	9.360290	21.940874	19.882891	2.399886e+01	51.396034
United States	1.439747	4.487700	22.960101	10.225201	13.082273	7.368130e+00	39.112749
India	0.150622	17.857780	2.541777	6.641372	13.282745	3.045590e-07	27.191551
Japan	1.570688	1.857906	8.745761	2.208639	2.456845	1.960433e+00	14.382993
Brazil	0.816372	2.854914	3.389111	5.481133	4.196472	6.765794e+00	12.541530

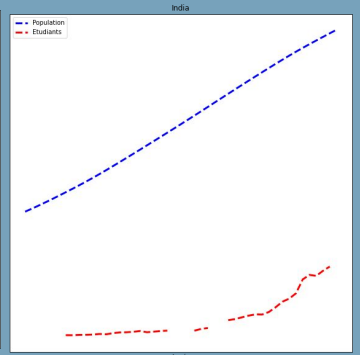
China versus USA

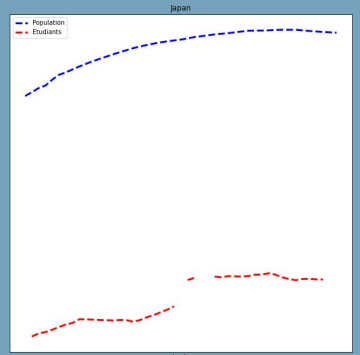


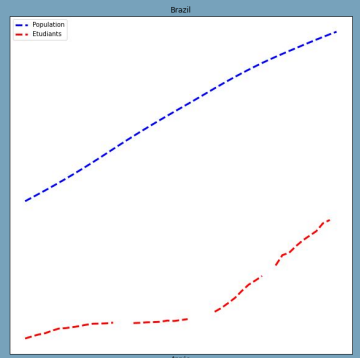
Recherche de corrélation











Population combinée avec le nombre d’étudiants à la faculté

Population
Etudiants

3. Conclusion sur l’implantation

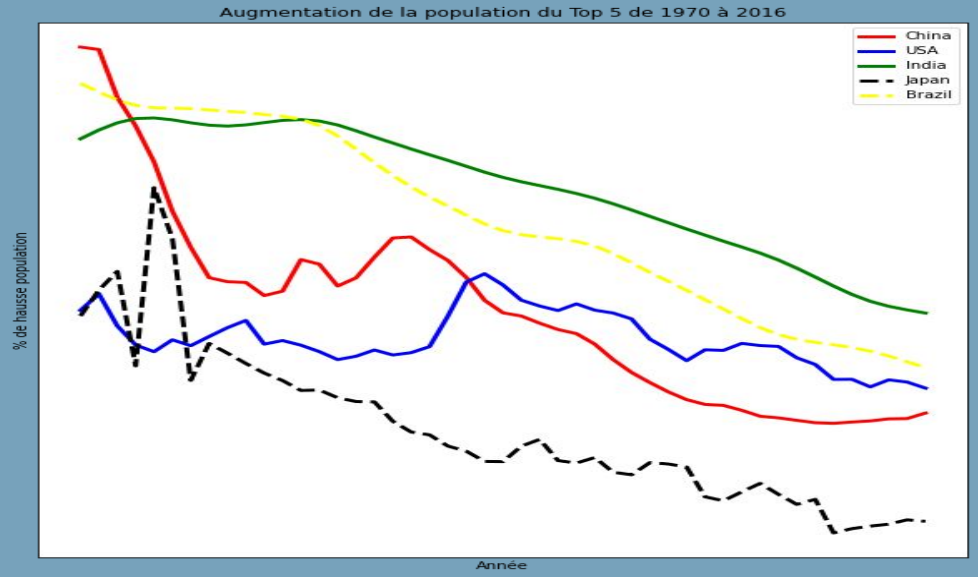
Top 5 des meilleures implantations

Les pays présentant le meilleur potentiel d’implantation sont par ordre d’importance :

1. China
2. USA
3. Inde
4. Japan
5. Brazil

3. Conclusion sur l'évolution

Projection utilisant la hausse de population



3. Conclusion sur le jeu de données

LA BANQUE MONDIALE

Grande crédibilité
Tous les pays du monde
Diversité des informations
Données propres et cohérentes

Faible complétude
Manque de fraîcheur
Manque données business
Meilleures données prospectives

Recommandations :

- Obtenir des jeux de données spécifiques pour le domaine du e-learning
- Croiser les résultats avec ces autres jeux de données
- Accueillir les retours d'academy pour accentuer des points de l'étude (scoring)

4. Axes d’amélioration de l’étude

- Vérifier les erreurs lexicales, d’irrégularités, outliers
- Améliorer la complétude des données (imputation aux lieux d’amputation)
- Utiliser la corrélation Pearson (np.corrcoef)
- Utiliser l'analyse en composantes principales : ACP
- Améliorer le code python et le simplifier

Questions et Réponses

(5 à 10 min)

Merci pour votre attention