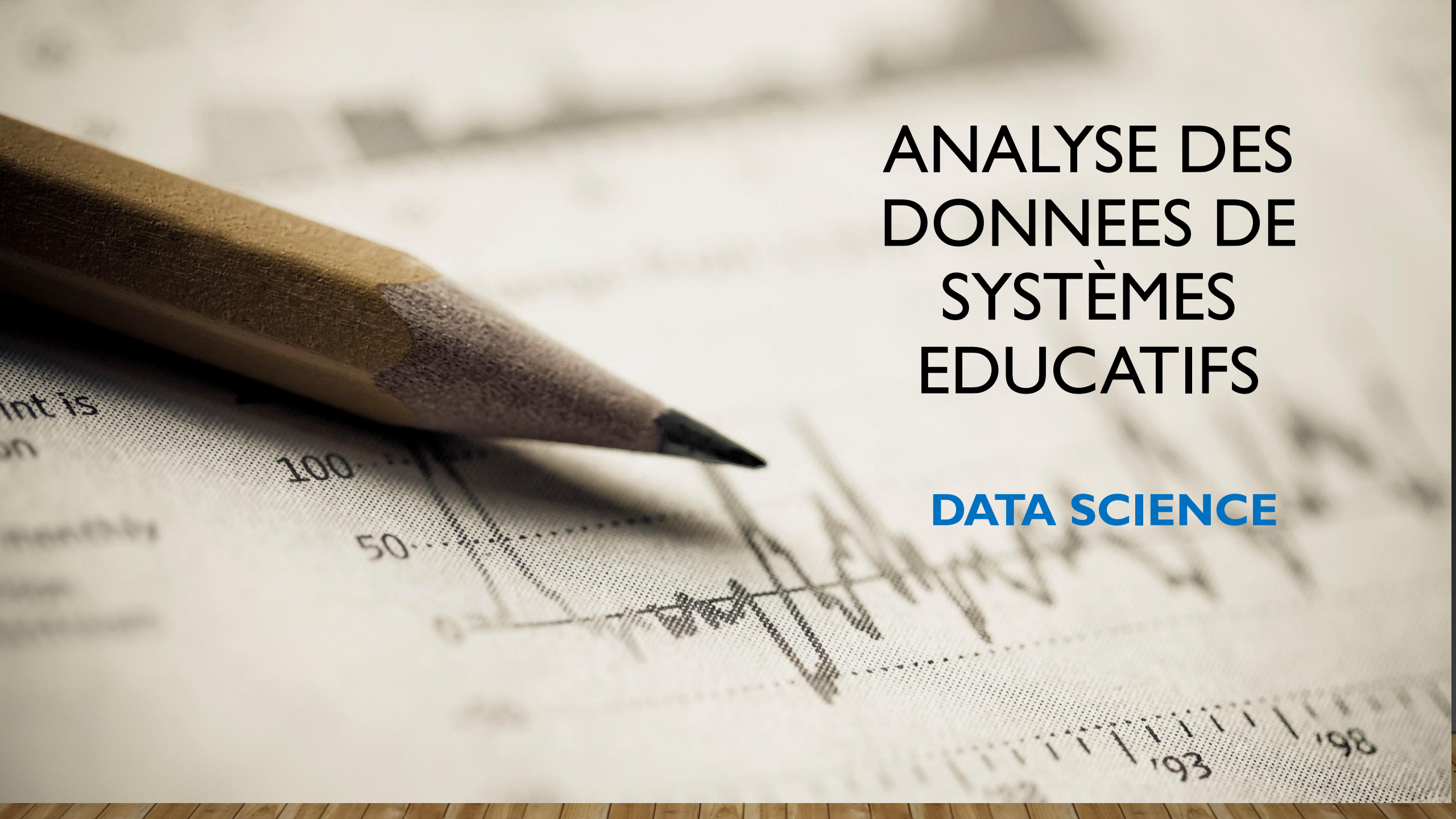


ANALYSE DES DONNEES DE SYSTÈMES EDUCATIFS

DATA SCIENCE

MÉTHODOLOGIE

I. PRESENTATION DU PROJET

II.VALIDATION DES DONNEES

III.ANALYSE DES DONNEES

IV.CONCLUSION

PRESENTATION DU PROJET

Contexte

- Start-up de formation en ligne
- Niveau lycée et université
- Extension d'activités dans d'autres pays

Objectifs

- Identifier pays à fort potentiel
- Analyser pour chaque pays l'évolution du potentiel
- Déterminer le(s) pays où opérer prioritairement

Contraintes

- BD de + 4.000 indicateurs
- Nombreuses données manquantes
- Probables doublons

VALIDATION DES DONNEES

Description du jeu de données

Data

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	1974	1975	...	2055	2060	2065	2070	2075
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
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
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.991138	59.36554	...	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
			Youth illiterate population													

Country

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group	WB-2 code	...	Government Accounting concept	IMF data dissemination standard	Latest population census	I hous s
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	NaN	2010	
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income	AF	...	Consolidated central government	General Data Dissemination System (GDDS)	1979	M Inc C S (M 20
2	AGO	Angola	Angola	People's Republic of Angola	AO	Angolan kwanza	April 2013 database update: Based on IMF data,...	Sub-Saharan Africa	Upper middle income	AO	...	Budgetary central government	General Data Dissemination System (GDDS)	1970	M Inc S (MIS)
3	ALB	Albania	Albania	Republic of Albania	AL	Albanian lek	NaN	Europe & Central Asia	Upper middle income	AL	...	Budgetary central government	General Data Dissemination System (GDDS)	2011	Demog and H S (20
4	AND	Andorra	Andorra	Principality of Andorra	AD	Euro	NaN	Europe & Central Asia	High income: nonOECD	AD	...	NaN	NaN	2011. Population figures compiled from adminis...	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
236	XKX	Kosovo	Kosovo	Republic of Kosovo	NaN	Euro	Kosovo became a World Bank member on June 29, ...	Europe & Central Asia	Lower middle income	KV	...	NaN	General Data Dissemination System (GDDS)	2011	
							Based on	Middle							

VALIDATION DES DONNEES

Description du jeu de données

Le jeu de données est un classeur Excel de plusieurs feuilles.

Nous utiliserons deux feuilles de notre jeu de données :

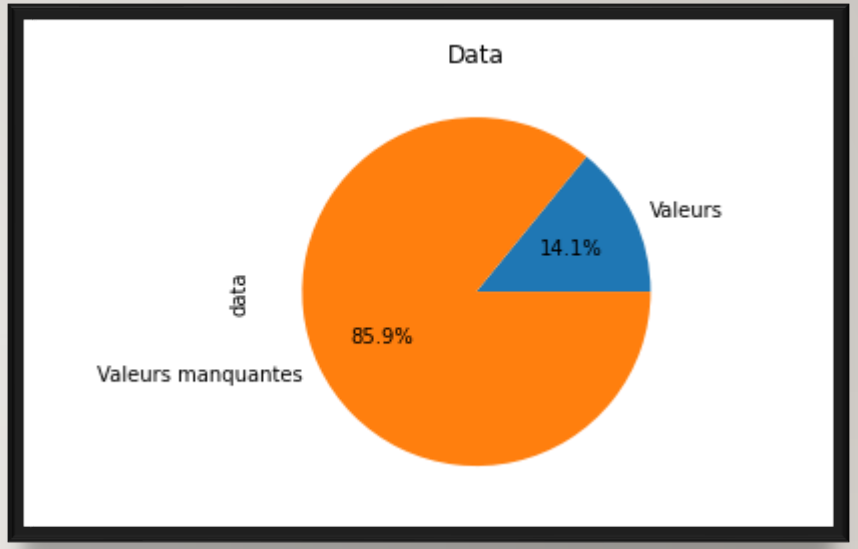
- ✓ "Data" : 886.930 lignes et 69 colonnes
- ✓ "Country" : 241 lignes et 31 colonnes

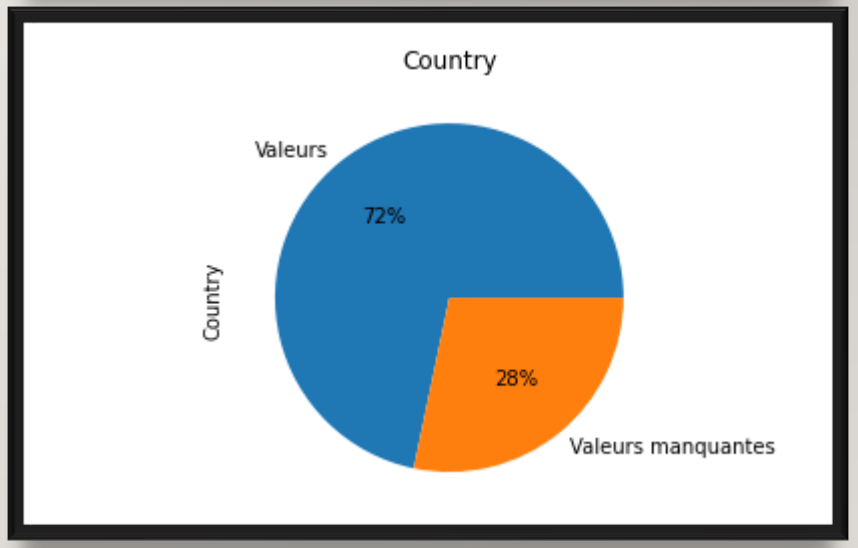
Toutefois la principale Data Frame a exploiter sera "Data"

VALIDATION DES DONNEES

Qualité du jeux de données

Data Frame	Nbre de données	Nbre de données manquantes
Data	8.629.921	52.568.249
Country	5.358	2.113





VALIDATION DES DONNÉES

Traitement des données

Elle va consister à :

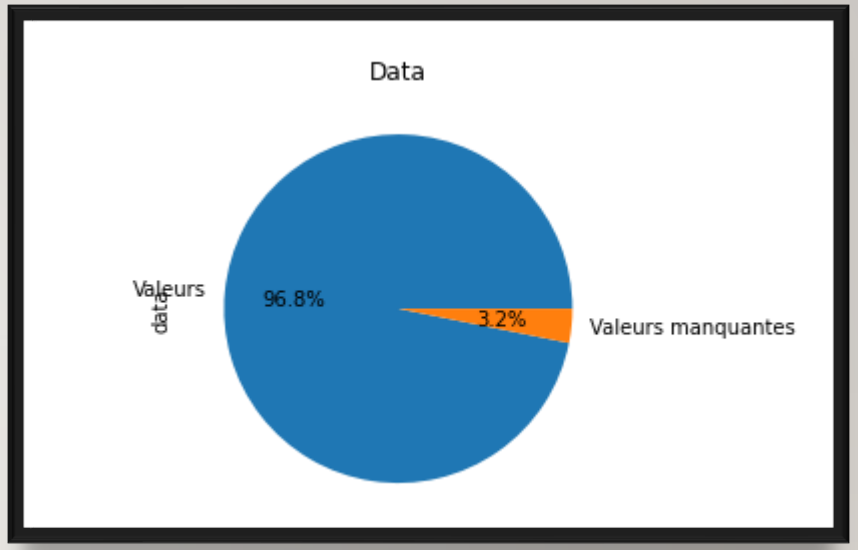
- Nettoyer le jeu de données
  - ✓ Les données manquantes
  - ✓ Les doublons probables
  - ✓ Les éventuels problèmes format des données sur chacune des colonnes
- Identifier les informations pertinentes :
  - ✓ Les colonnes
  - ✓ Les lignes

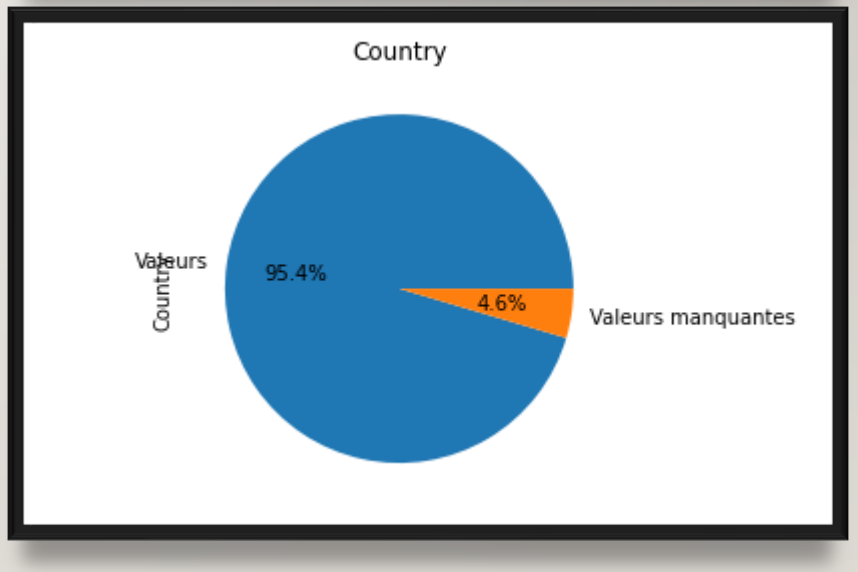
VALIDATION DES DONNEES

Traitement des données

- Nettoyage du jeu de données

Data Frame	Nbre de données	Nbre de données manquantes
Data	1.133.698	36.962
Country	3.676	176





VALIDATION DES DONNÉES

Traitement des données

- Identifier les informations pertinentes :
  - Colonne :
    - ✓ « Data » : Identification de six (6) colonnes pertinentes ;
    - ✓ « Country » : Identification de deux (2) colonnes pertinentes ;
    - ✓ Fusionner « Data » et « Country »

VALIDATION DES DONNÉES

Traitement des données

- Identifier les informations pertinentes :
  - Lignes (indicateurs sur l'éducation)
    - ✓ 'Gross enrolment ratio, upper secondary, both sexes (%)',
    - ✓ 'Gross enrolment ratio, tertiary, both sexes (%)',
    - ✓ 'Internet users (per 100 people)',
    - ✓ 'Population of the official age for upper secondary education, both sexes (number)',
    - ✓ 'Population of the official age for tertiary education, both sexes (number)'

	lignes	colonnes
Data	557	8

VALIDATION DES DONNÉES

La base de données pour l'analyse

	Region	Country Name	Indicator Name	2010	2011	2012	2013	2015
0	Latin America & Caribbean	Aruba	Gross enrolment ratio, tertiary, both sexes (%)	37.3577	38.737621	37.755692	16.195829	15.273780
1	Latin America & Caribbean	Aruba	Internet users (per 100 people)	62.0000	69.000000	74.000000	78.900000	88.661227
2	Latin America & Caribbean	Aruba	Population of the official age for tertiary ed...	6237.0000	6464.000000	6942.000000	7292.000000	7634.000000
3	Latin America & Caribbean	Aruba	Population of the official age for upper secon...	4623.0000	4594.000000	4553.000000	4529.000000	4492.000000
4	Sub-Saharan Africa	Angola	Internet users (per 100 people)	2.8000	3.100000	6.500000	8.900000	12.400000

ANALYSE DES DONNEES

- Indicateurs Statistiques
- Tri par Indicateurs sur l’éducation
- Courbe d’évolution

ANALYSE DES DONNEES

Indicateurs Statistiques

Nous allons :

- Grouper les pays par zone géographique
- Calculer la Moyenne, la Médiane et l'Ecart-type pour chaque zone

ANALYSE DES DONNEES

Indicateurs Statistiques

Nous identifions deux (2) régions à fort potentiel :

- ✓ Europe et Asie Centrale
- ✓ Asie de l'est et Pacifique

ANALYSE DES DONNEES

Tri sur les Indicateurs

Nous allons effectuer :

- Tri décroissant sur chacun des cinq (5) indicateurs ;
- Liste des dix (10) premiers pays pour chaque tri
- Identification des cinq (5) pays qui reviennent régulièrement

ANALYSE DES DONNEES

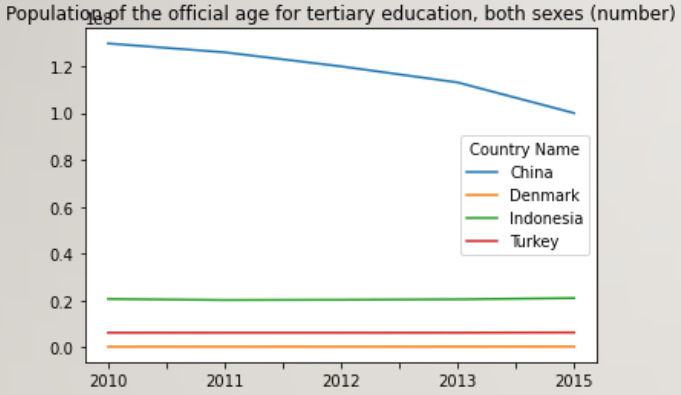
Tri sur les Indicateurs

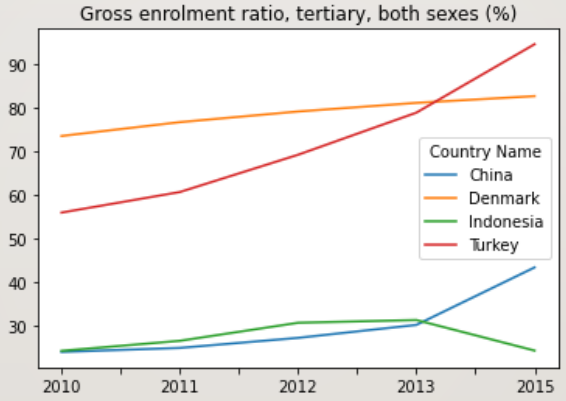
La liste des cinq (5) pays qui en résulte :

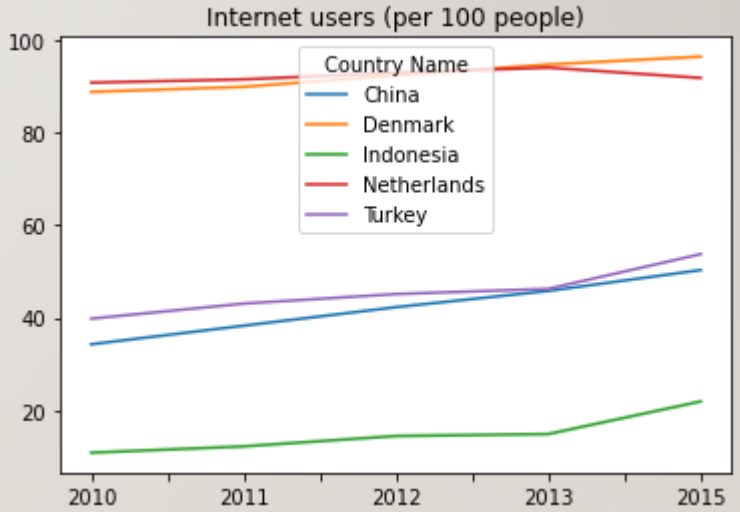
- ✓ Turquie (3/5) ;
- ✓ Danemark (3/5) ;
- ✓ Hollande (2/5) ;
- ✓ Chine (2/5) ;
- ✓ Indonésie (2/5)

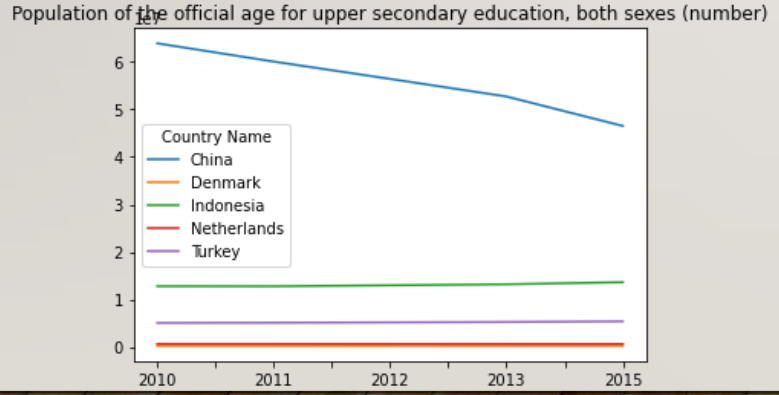
ANALYSE DES DONNEES

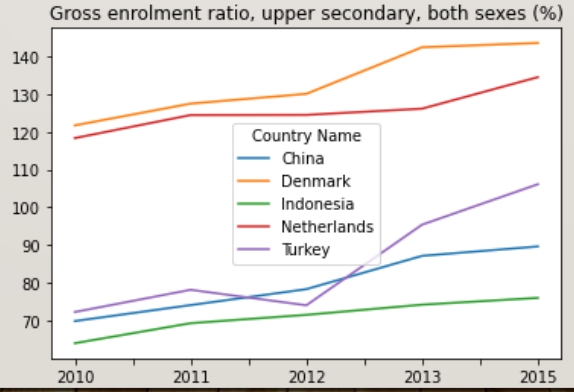
Courbes d'évolution











CONCLUSION

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

Nous avons identifié :

- ***'Asie de l'est et Pacifique'*** : l'Indonésie et la Chine;
- ***'Europe et Asie Centrale'*** : la Turquie, le Danemark, la Hollande.

CONCLUSION

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

Nous constatons :

- ***La hollande,*** n' a de données que sur 3/5 indicateurs; également une décroissance à partir de l'année 2013 pour l'un des indicateurs ;

- ***La Chine,*** connait une décroissance sur deux(2) de ces courbes d'évolution ;

- ***L'Indonésie***, une croissance sur les cinq (5) indicateurs; toutefois une décroissance sur l'un des indicateurs à partir de l'année 2013;

- ***Le Danemark et la Turquie,*** ont une évolution croissante sur tous les indicateurs sur l'éducation.

CONCLUSION

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

Après une visualisation de l’évolution du potentiel de ces différents pays :

- **le Danemark et la Turquie** restent d’excellents potentiels pour le service d’Academy, qu’il faudrait mettre en priorité ;
- **L’Indonésie** parait également comme un potentiel uniquement pour le niveau du lycée.

MERCI

OPENCLASSROOM