24/07/23

Analyse des données de systèmes éducatifs

Armand FAUGERE LinkedIn
armand-faugere@live.fr

Sommaire

1) Cadrage du projet
2) Données d'entrée
3) Démarche de traitement des données
4) Démarche d'analyse des données
5) Synthèse des résultats
6) Conclusion

1) Cadrage du projet

1) Cadrage du projet

❑ **Contexte :** projet d'extension à l'international d' ACADEMY qui propose des contenus de formation en ligne pour un public de niveau Lycée et Université

❑ **But :**

- Déterminer si les données sur l'éducation de la banque mondiale peuvent informer le projet

❑ **Objectifs :**

- Réaliser une analyse pré-exploratoire du jeu de donnée sur la banque mondiale
- Valider la qualité du jeu de données
- Décrire les informations contenues dans le jeu de données
- Sélectionner les informations pertinentes
- Déterminer des ordres de grandeur des indicateurs

2) Données d’entrée

2) Données d'entrée

❑ Les jeux de données

THE WORLD BANK | Data Catalog
IBRD • IDA

➔ Statistiques sur l'éduction

https://datacatalog.worldbank.org/search/dataset/0038480

- Base de données annuelle (enquêtes, sondages…) ➔ 22/01/23
- Plus de 4000 indicateurs (accès éduction, enseignement, population dépenses…)
- Couverture de plusieurs années jusque 1970 – 2050
- 5 fichiers de données en format csv

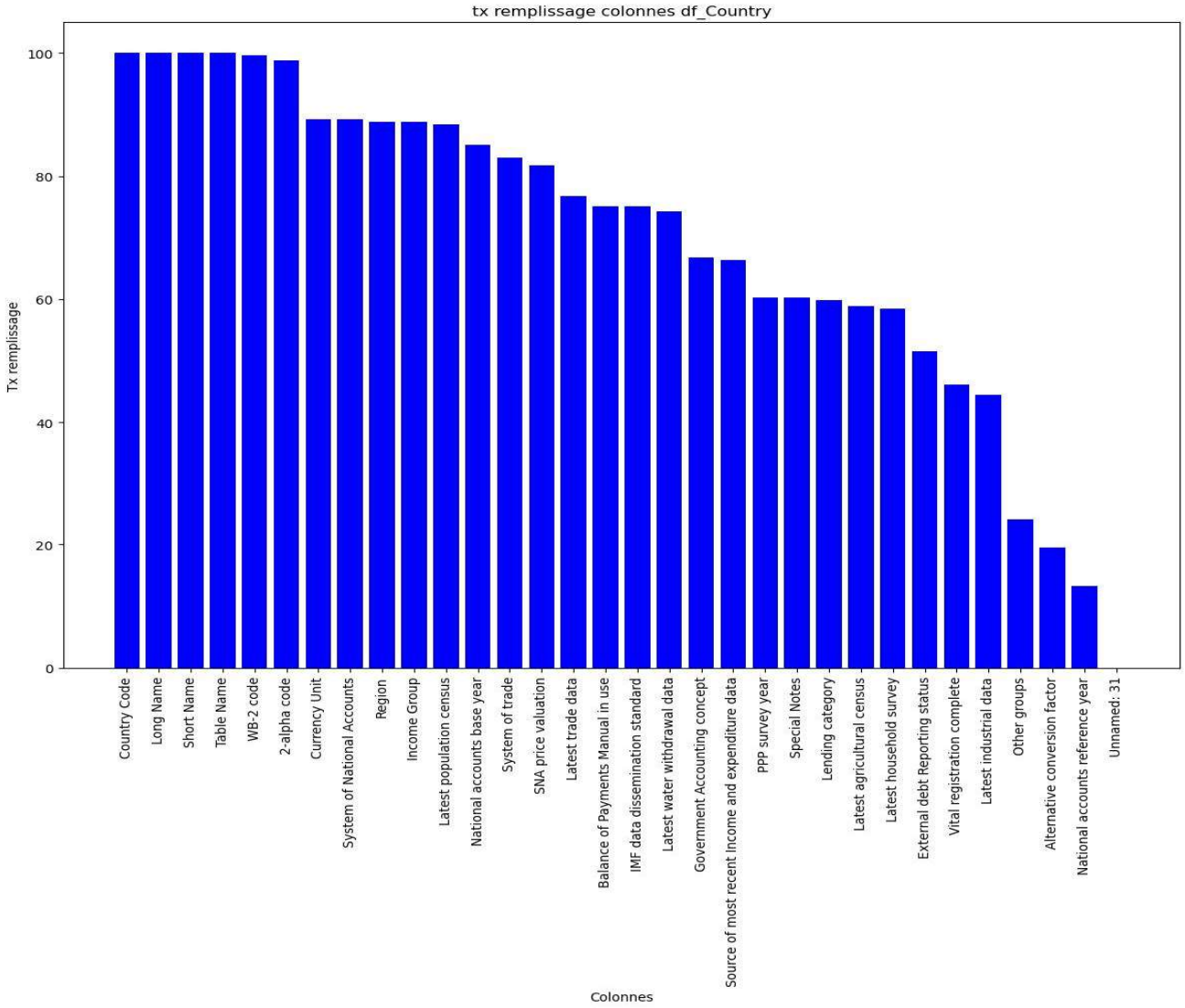
2) Données d'entrée

❑ **Le du jeu de données**

➔ **EdStatsCountry.csv**

- 32 colonnes ➔ type de données
- 241 lignes ➔ pays

❑ **Graphique taux de remplissage des données en %**



Il y a assez peu de colonnes complètes

2) Données d'entrée

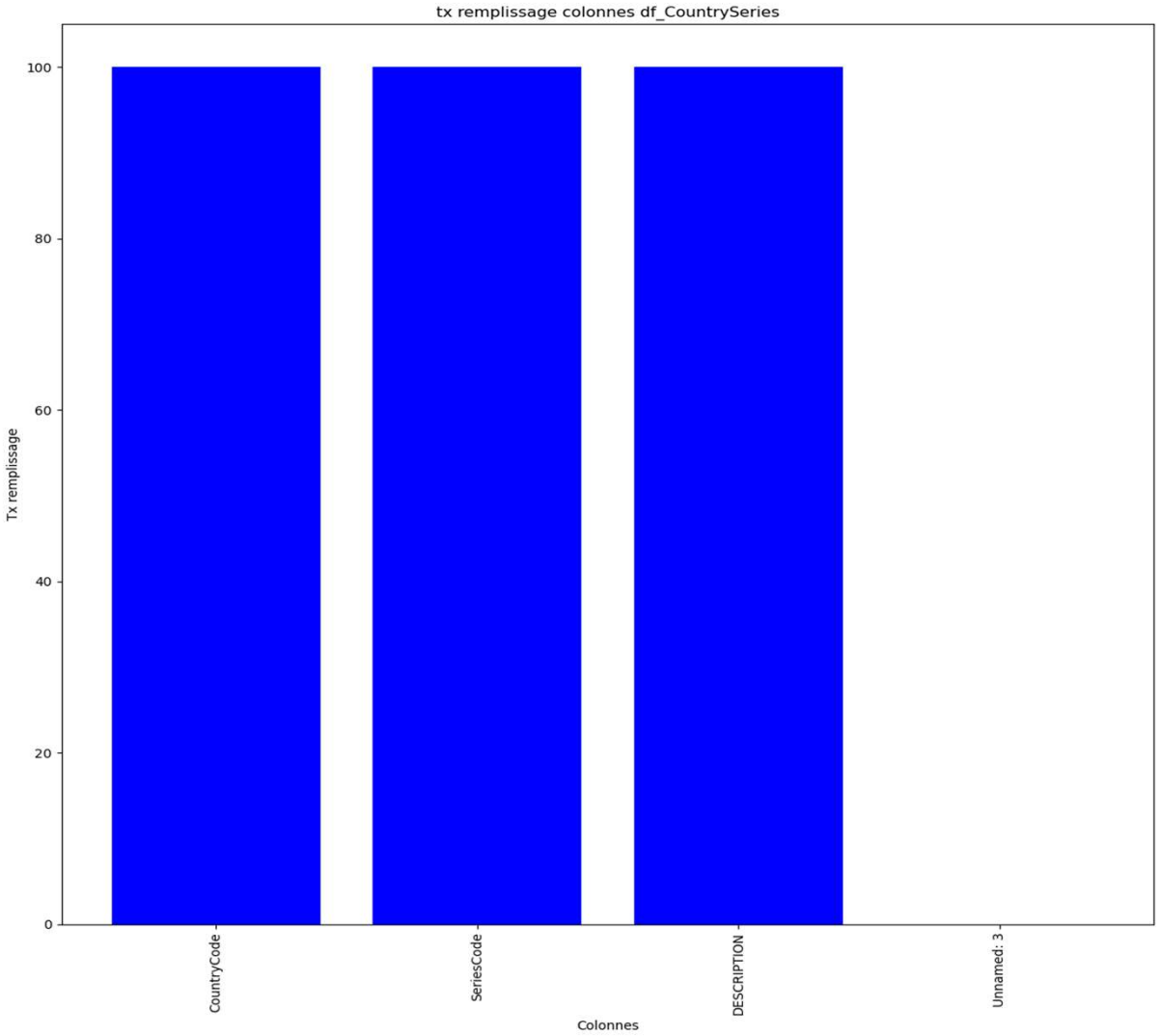
❑ **Le jeu de données**

➔ **EdStatsCountry-Series.csv**

- 4 colonnes ➔ type de données

- 613 lignes ➔ Codes pays et indicateurs

❑ **Graphique taux de remplissage des données en %**



Les colonnes nécessaires sont à 100%

2) Données d'entrée

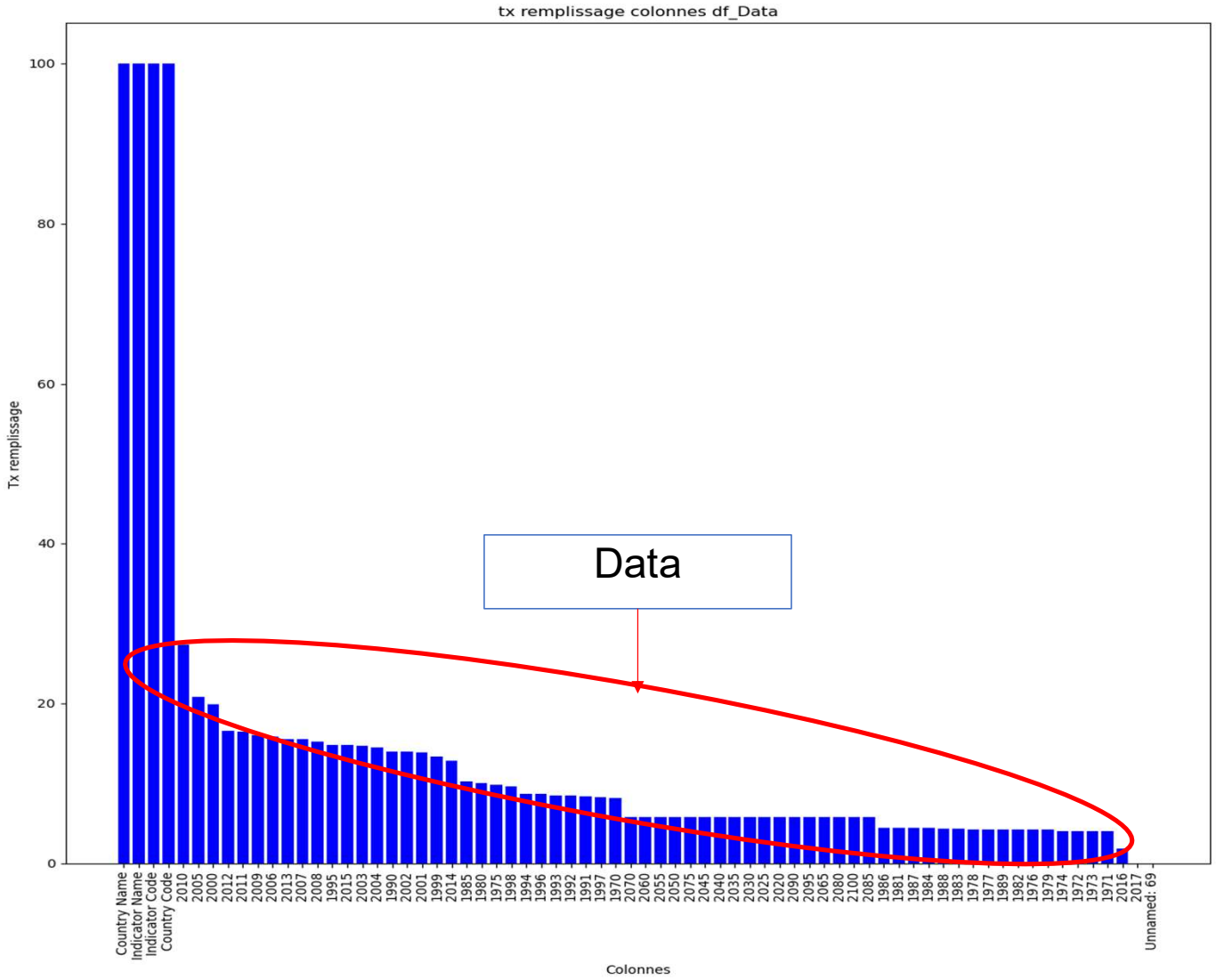
❑ **Le jeu de données**

➔ **EdStatsData.csv**

- 70 colonnes ➔ type de données + années avec valeur des indicateurs

- 886930 lignes ➔ Pays et indicateurs

❑ **Graphique taux de remplissage des données en %**



Il y a assez peu de colonnes complètes

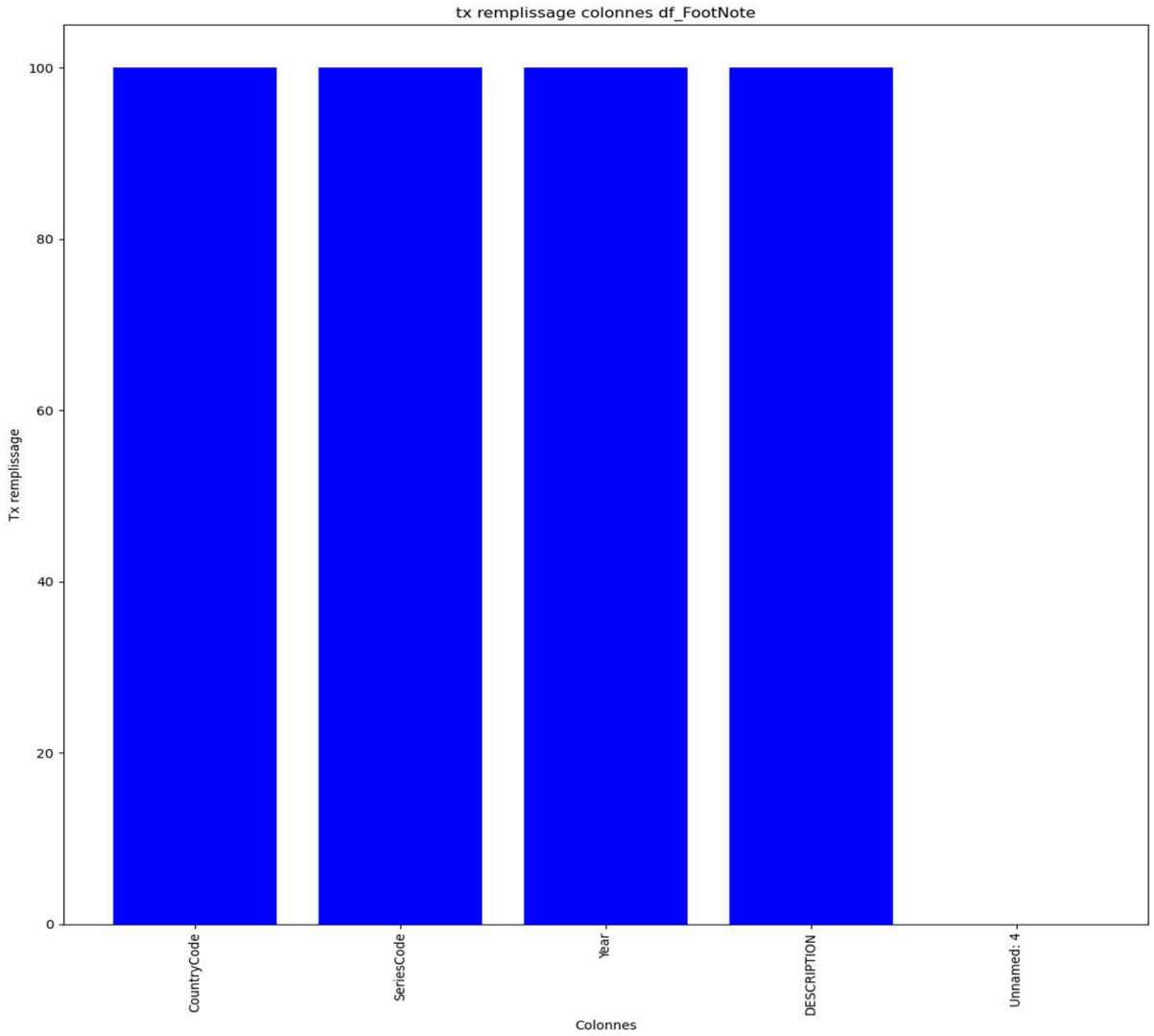
2) Données d'entrée

❑ **Le du jeu de données**

➔ **EdStatsFootNote.csv**

- 5 colonnes ➔ type de données
- 643638 lignes ➔ Code pays et indicateurs

❑ **Graphique taux de remplissage des données en %**



Les colonnes nécessaires sont à 100%

2) Données d'entrée

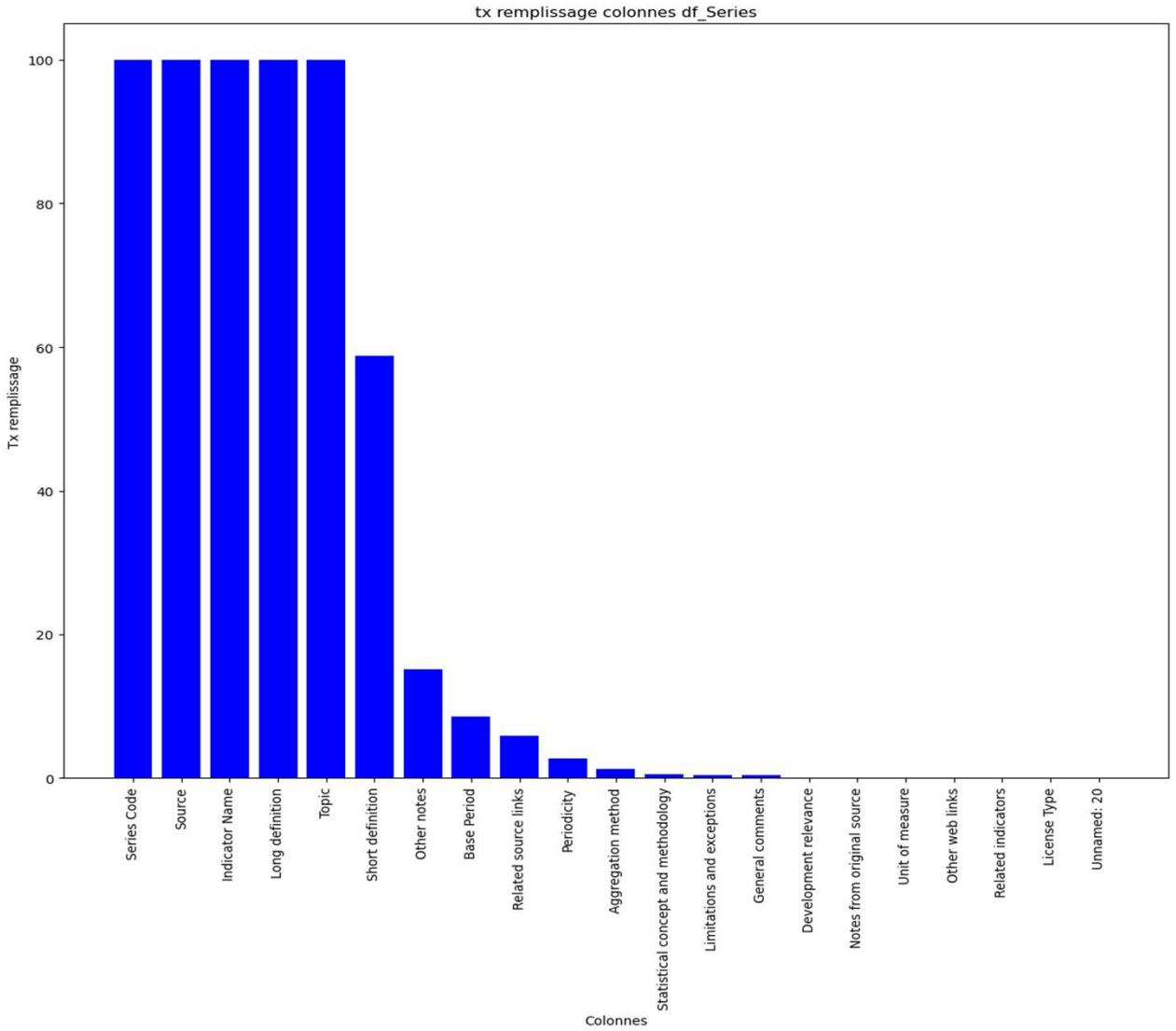
❑ **Le du jeu de données**

➔ **EdStatsSeries.csv**

- 21 colonnes ➔ type de données

- 3665 lignes ➔ Codes et Type d'indicateur

❑ **Graphique taux de remplissage des données en %**



Les colonnes intéressantes sont à 100%

2) Données d’entrée

Les données brutes sont inexploitables.

Un travail important de traitement de données a été réalisé pour tirer les éléments pertinents.

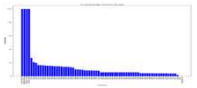
3) Démarche de traitement des données

3) Démarche de traitement des données

1

Traitement des jeux de données pour un jeu de données exploitable

Jupiter Notebook, Python, *Pandas, Numpy, Matplotlib, Seaborn*



Taux de remplissage Colonne

Analyse du jeu de données

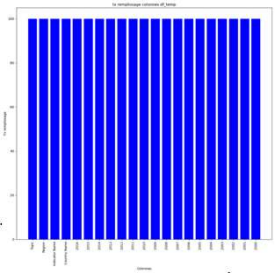
Analyse du contenu des colonnes pertinentes (Topic, Indicator_Name...)

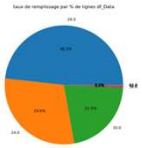
Exploration des variables

*Fonctions de suppression **des colonnes** et **lignes** en fonction du % de remplissage*

Traitement des valeurs nulles NAN (boucles itératives)

Taux de remplissage Colonne à 100%





Taux de remplissage % de lignes

Suppression des colonnes selon % de remplissage

Fusion des jeux de données selon critères spécifiques (Indicateurs, séries, puis country, Région) ➔ un seul dataframe

Suppression des lignes selon tx de remplissage

Suppression des colonnes à tx de remplissage 0%

Suppression des Topics indicateurs hors sujets

Suppression des indicateurs hors sujets

Sélection des indicateurs par fonctions REGEX ➔ 155 indicateurs

Choix de 13 Indicateurs

Remplacement des dernières valeurs manquantes en remplaçant sur une année précédente ou suivante

3) Démarche de traitement des données

Indicateurs retenus pour la suite de l'analyse :

Données économiques
➔ GDP per capita (current US$)
➔ GNI (current US$)

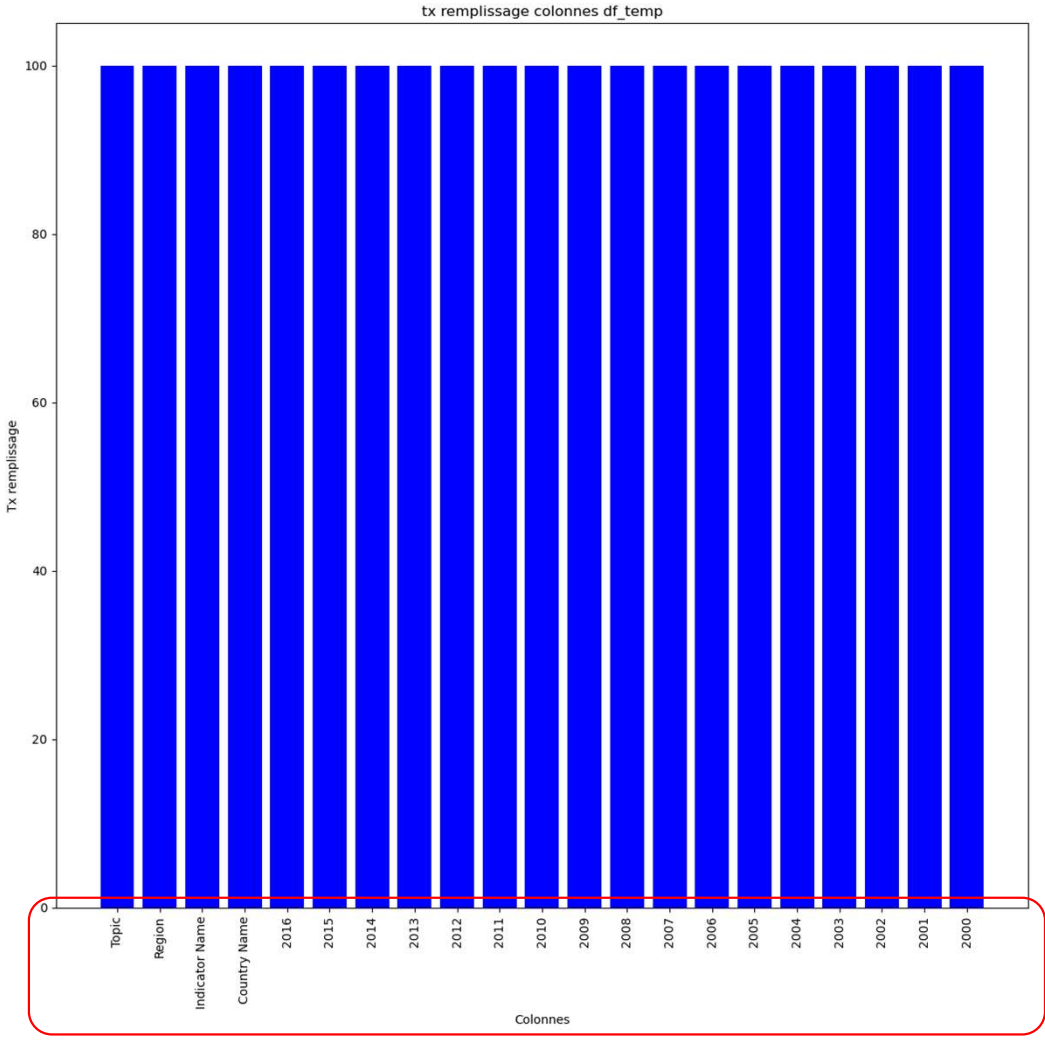
Marché cible Population
➔ Population, ages 15-24, total

Moyens de communications
➔ Internet users (per 100 people)

Marché privée actuel
➔ Enrolment in secondary education, private institutions, both sexes (number)
➔ Enrolment in upper secondary education, private institutions, both sexes (number)
➔ Enrolment in post-secondary non-tertiary education, private institutions, both sexes (number)
➔ Percentage of enrolment in secondary education in private institutions (%)
➔ Percentage of enrolment in upper secondary education in private institutions (%)
➔ Percentage of enrolment in post-secondary non-tertiary education in private institutions (%)
➔ Percentage of enrolment in tertiary education in private institutions (%)

Marché global
➔ Enrolment in post-secondary non-tertiary education, both sexes (number)
➔ Enrolment in tertiary education, all programmes, both sexes (number)

Ratio élève enseignant
➔ Pupil-teacher ratio in tertiary education (headcount basis)
➔ Pupil-teacher ratio in upper seconda



3) Démarche de traitement des données

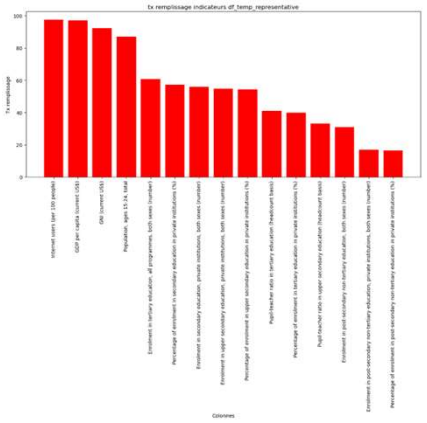
2

Traitement du jeu de données pour extraire les indicateurs et les pays/régions exploitables

Jupiter Notebook, Python, *Pandas, Numpy, Matplotlib, Seaborn*

208 pays et ils n'ont pas tous les indicateurs renseignés

Analyse des 13 indicateurs en termes de **représentativité** (données suffisantes pour comparer les pays et région)



Taux de remplissage Indicateur

*Fonctions de suppression **des colonnes** et **lignes** en fonction du % de remplissage*

Traitement des valeurs nulles NAN (boucles itératives)

Check des valeurs nulles pour chaque indicateur

Suppression des pays non intéressant d'un point de vue business (taille, intérêt..)

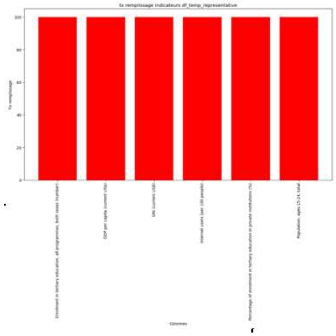
Suppression des pays avec trop peu de données pour lancer un projet d'investissement

Sélection de pays avec des filtres type moyenne, 1er quartile

Suppression des indicateurs pas assez représentatifs

Choix de 6 Indicateurs

Taux de remplissage Indictaur à 100%



3) Démarche de traitement des données

- ❑ 43 pays
- ❑ 5 Régions géographiques
- ❑ 6 indicateurs

1) MARCHE

Population, ages 15-24, total
➔ Très important, tranche d'age à démarcher

Enrolment in tertiary education, all programmes, both sexes (number)
➔ Très important pour démontrer le nombre d'étudiant dans le cycle le plus haut

Percentage of enrolment in tertiary education in private institutions (%)
➔ Très important pour démontrer la pénétration du marché privée sur le haut niveau d'étude

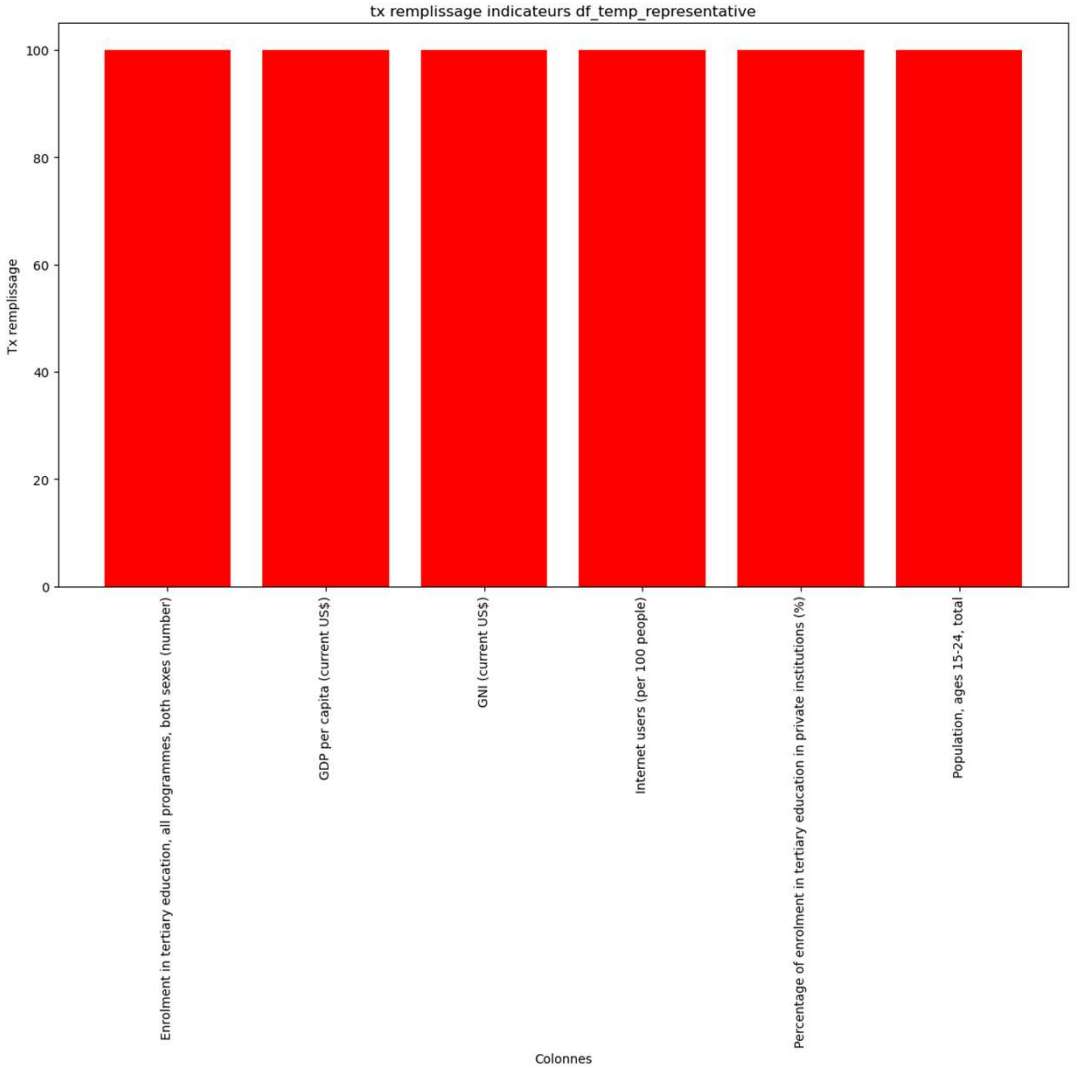
2) ECONOMIE

GDP per capita (current US$)
➔ important pour démontrer les richesses et leur répartion sur les habitants

GNI (current US$)
➔ Très important pour garantir les moyens de suivre la formation et le potentiel d'activité ensuite

3) INFRASTRUCTURES

Internet users (per 100 people)
➔ Primordial pour déployer le business



4) Démarche d’analyse des données

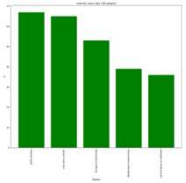
4) Démarche d'analyse des données

3

Exploration des indicateurs (Moyenne des années)

Par Région & Par Pays

Etablissement d'un Score de classification

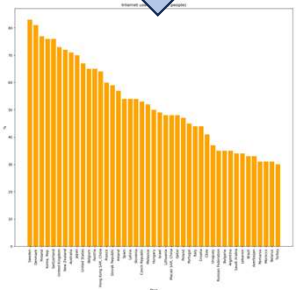


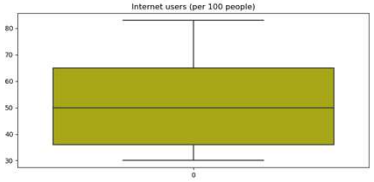
Jupiter Notebook, Python, *Pandas, Numpy, Matplotlib, Seaborn, sklearn*

Pas de critère de classification évidents

Analyse des 6 indicateurs par Région

Analyse des 6 indicateurs par pays





Création d'un Score de classification

Normalisation des données (mise à l'échelle)

Pondération de certains indicateurs

Création d'un score pour permettre un classement sur un seul indicateur

Score de classification

4) Démarche d'analyse des données

Top Région/indicateur

Internet users (per 100 people)

North America	67.0
East Asia & Pacific	65.0
Europe & Central Asia	53.0
Middle East & North Africa	39.0
Latin America & Caribbean	36.0

GNI (current US$)

North America	1.456132e+13
East Asia & Pacific	1.082453e+12
Europe & Central Asia	5.175165e+11
Latin America & Caribbean	4.999893e+11
Middle East & North Africa	1.757730e+11

GDP per capita (current US$)

North America	46989.0
East Asia & Pacific	31219.0
Europe & Central Asia	24270.0
Middle East & North Africa	22916.0
Latin America & Caribbean	9110.0

Pas pertinent

Approche par Moyenne des Moyennes

Percentage of enrolment in tertiary education in private institutions (%)

Latin America & Caribbean	47.0
East Asia & Pacific	40.0
Middle East & North Africa	35.0
North America	26.0
Europe & Central Asia	22.0

Enrolment in tertiary education, all programmes, both sexes (number)

North America	18075424.0
Latin America & Caribbean	2277520.0
East Asia & Pacific	1380309.0
Europe & Central Asia	1050862.0
Middle East & North Africa	378863.0

Population, ages 15-24, total

North America	43485037.0
Latin America & Caribbean	11104224.0
East Asia & Pacific	4256792.0
Europe & Central Asia	3306166.0
Middle East & North Africa	2590067.0

4) Démarche d'analyse des données

Top 5 Pays/indicateur

Internet users (per 100 people)

Sweden	83.0
Denmark	81.0
Finland	77.0
Korea, Rep.	76.0
Switzerland	76.0

GNI (current US$)

United States	1.456132e+13
Japan	5.067986e+12
United Kingdom	2.470854e+12
France	2.380305e+12
Italy	1.876628e+12

GDP per capita (current US$)

Switzerland	65759.0
Qatar	60779.0
Denmark	51300.0
Ireland	49393.0
Sweden	47110.0

Percentage of enrolment in tertiary education in private institutions (%)

United Kingdom	100.0
Latvia	95.0
Israel	83.0
Korea, Rep.	80.0
Chile	79.0

Enrolment in tertiary education, all programmes, both sexes (number)

United States	18075424.0
Russian Federation	8116286.0
Brazil	5756329.0
Japan	3938675.0
Turkey	3259860.0

Population, ages 15-24, total

United States	43485037.0
Brazil	34369769.0
Russian Federation	20868804.0
Turkey	13703175.0
Japan	13517999.0

4) Démarche d'analyse des données

Type	Indicateur	Facteur multiplicateur
MARCHE	Population, ages 15-24, total	**2**
	Enrolment in tertiary education, all programmes, both sexes (number)	**2**
	Percentage of enrolment in tertiary education in private institutions (%)	**2**
ECONOMIE	GDP per capita (current US$)	**1**
	GNI (current US$)	**2**
INFRA	Internet users (per 100 people)	**3**

Score =
(Population, ages 15-24, total X 2)
+ (Enrolment in tertiary education, all programmes, both sexes (number) X 2)
+ (Percentage of enrolment in tertiary education in private institutions (%) X 2)
+ (GDP per capita (current US$) X 1)
+ (GNI (current US$) X 2)
+ (Internet users (per 100 people) X 3)

4) Synthèse des résultats

5) Synthèse des résultats

❑ **Tops countries**
➔ **Score des pays > 0**

Country Name	Region	SCORE
United States	North America	35.183532
Japan	East Asia & Pacific	14.766877
United Kingdom	Europe & Central Asia	12.368512
Korea, Rep.	East Asia & Pacific	9.530565
Brazil	Latin America & Caribbean	8.363081
Switzerland	Europe & Central Asia	4.010730
France	Europe & Central Asia	3.677770
Sweden	Europe & Central Asia	3.597391
Belgium	Europe & Central Asia	3.184943
Russian Federation	Europe & Central Asia	2.857293
Denmark	Europe & Central Asia	2.663186
Finland	Europe & Central Asia	2.548444
Australia	East Asia & Pacific	2.190857
Latvia	Europe & Central Asia	1.512803
Israel	Middle East & North Africa	1.152941
New Zealand	East Asia & Pacific	0.260578
Austria	Europe & Central Asia	0.202619

5) Synthèse des résultats

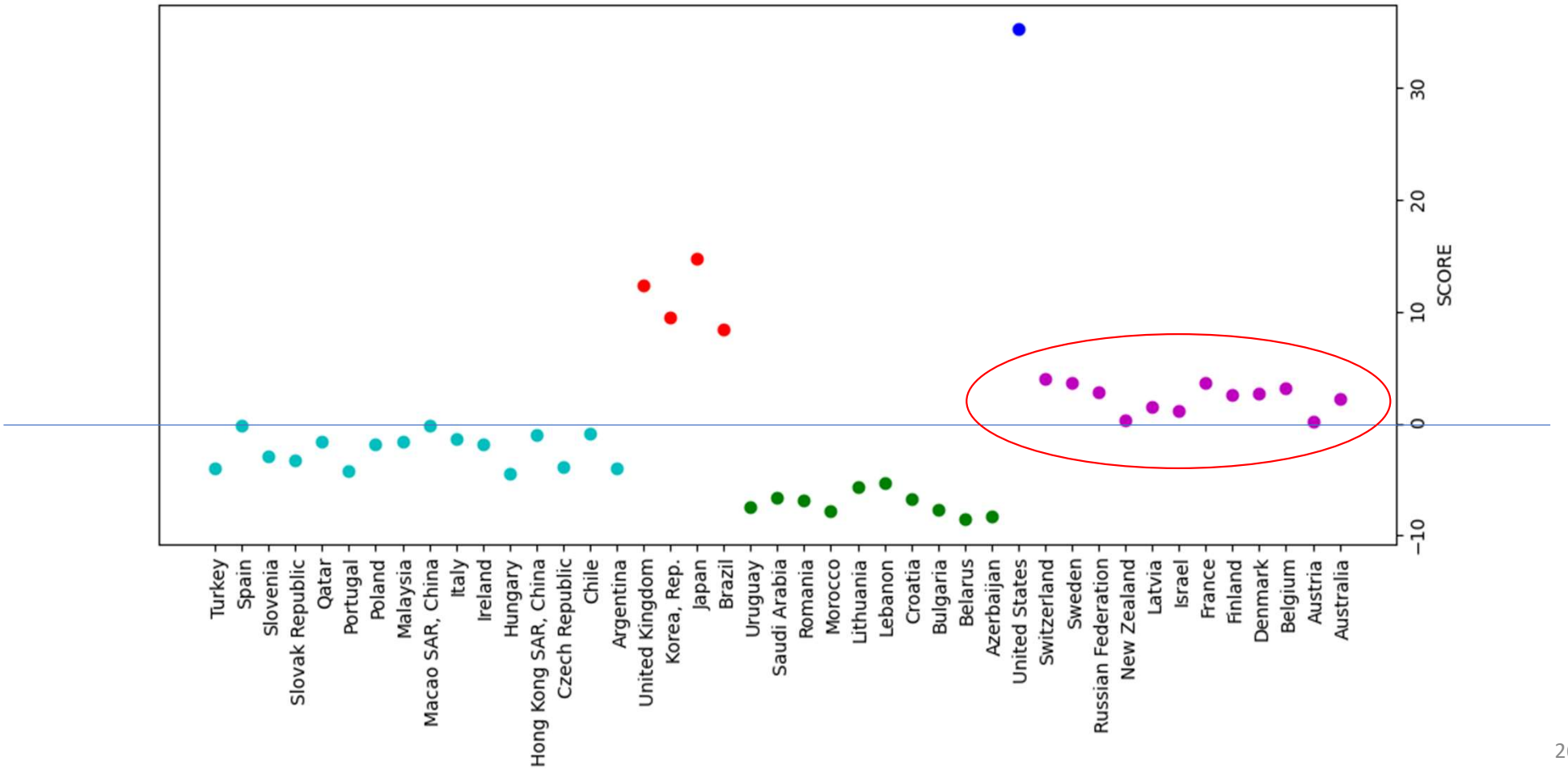
❑ **Tops région**

➔ **Somme des scores des pays**

Région	SCORE
Europe & Central Asia	36.623693
North America	35.183532
East Asia & Pacific	26.748877
Latin America & Caribbean	8.363081
Middle East & North Africa	1.152941

5) Synthèse des résultats

❑ **Classification de tous les pays (score proche)**



5) Synthèse des résultats

❑ **La région préférentielle pour investir est <u>l'Europe et Asie centrale</u>**

❑ **Les pays préférentiels sont :**

United Kingdom*
Switzerland
France
Sweden
Belgium
Russian Federation
Denmark
Finland
Latvia
Austria

En investissant dans ces pays à fort potentiel, mais aussi proches en termes de score, les **risques** et les **investissements** seront mieux maitrisés

5) Conclusion

Conclusion

- **Les jeux de données de la banque mondiale sont très complexes, avec énormément de données dont beaucoup sont manquantes, mais une analyse fine a pu être réalisée**

- **Les Top Régions et les Tops pays ont pu être identifiés avec une approche méthodique et structurée**

- **Les résultats de l'analyse peuvent donc être utilisés dans le cadre du projet d'extension à l'international**

Merci

- Armand FAUGERE
- armand-faugere@live.fr