

Faculté des Sciences et Ingénierie - Sorbonne Université

Master Informatique parcours - DAC



**DALAS - Data science, Learning And ApplicationS**

## Rapport de projet

---

# L'intelligence est elle innée ou acquise ?

---

Réalisé par :

Ahmed Abdelaziz MOKEDDEM

Faten Racha SAID

Supervisé par :

Laure SOULIER

Mai 2024

# Table des matières

---

<b>Introduction</b>	<b>1</b>
<b>1 Cadrage du projet</b>	<b>2</b>
1.1 Contexte . . . . .	2
1.1.1 Perspective innéiste . . . . .	2
1.1.2 Perspective empiriste . . . . .	2
1.2 Environnement de travail . . . . .	3
1.2.1 Langage de programmation et éditeur . . . . .	3
1.2.2 Collecte et structuration des données . . . . .	3
1.2.3 Visualisation des données . . . . .	4
1.2.4 Analyse des données et modélisation . . . . .	4
1.2.5 Interface utilisateur et déploiement . . . . .	4
1.3 Acquisition des données . . . . .	5
<b>2 Analyse de données</b>	<b>9</b>
2.1 Analyse préliminaire . . . . .	9
2.1.1 Distribution des variables . . . . .	9
2.1.2 Valeurs renseignées . . . . .	15
2.2 Analyse Exploratoire des Données (EDA) . . . . .	16
2.2.1 Axe géoéconomique . . . . .	16
2.2.2 Axe historique . . . . .	17
2.2.3 Axe socio-économique . . . . .	20
2.2.4 Axe culturel et éducatif . . . . .	24
2.3 Analyse en composantes principales (ACP) . . . . .	28
2.3.1 Analyse des résultats . . . . .	30
<b>3 Implémentations et expérimentations</b>	<b>33</b>
3.1 Modèles de Classification . . . . .	33
3.1.1 SVM . . . . .	33
3.1.2 Régression logistique . . . . .	33
3.1.3 XGBoost . . . . .	34
3.1.4 Évaluation des Modèles . . . . .	34
3.1.5 Analyse des Résultats . . . . .	34
3.2 Modèles de Clustering . . . . .	35

3.2.1	K-means . . . . .	35
3.2.2	Clustering Agglomératif . . . . .	38
3.3	Modèles de Régression Linéaire . . . . .	39
3.3.1	Régression linéaire . . . . .	39
3.3.2	Réseau de neurones . . . . .	40
3.3.3	Comparaison des résultats . . . . .	42
3.4	Modèles de Séries Temporelles . . . . .	42
3.4.1	ARIMA . . . . .	42
3.4.2	Prophet . . . . .	48
<b>4</b>	<b>Mise en production de la solution</b>	<b>50</b>
4.1	Présentation du tableau de bord . . . . .	50
4.1.1	Home : Analyse exploratoire des données . . . . .	50
4.1.2	Regression : Prédiction du QI . . . . .	51
4.1.3	Clustering : Regroupement des pays en clusters . . . . .	52
4.1.4	Time Series : Prédiction de l'indice évolution dans le temps . . . . .	53
4.2	Déploiement . . . . .	53
4.2.1	Utilisation de Docker pour la Conteneurisation . . . . .	53
4.2.2	Construction de l'Image Docker et Exécution du Conteneur . . . . .	54
<b>Conclusion</b>		<b>55</b>

# Introduction

---

L'intelligence, souvent définie comme la capacité d'apprendre, de comprendre, de raisonner et de s'adapter à de nouvelles situations, est un concept complexe et multidimensionnel. Elle englobe une variété de compétences cognitives telles que la mémoire, la résolution de problèmes, la créativité et la pensée critique. Le QI est un indicateur quantitatif de l'intelligence humaine, basé sur des tests standardisés qui évaluent diverses capacités cognitives. Bien que controversé et limité dans sa capacité à saisir toute la complexité de l'intelligence humaine, le QI reste une mesure largement utilisée et acceptée dans les études psychométriques.

Est-elle acquise ou innée ? Cette question, qui a longtemps alimenté les débats parmi les scientifiques et les philosophes, demeure au cœur de nombreuses recherches contemporaines. Ce projet, dans le cadre de l'UE DALAS, se propose d'apporter une perspective nouvelle et éclairée sur ce sujet complexe en exploitant les outils et techniques modernes de la data science.

Dans ce rapport, nous détaillerons les différentes étapes et méthodologies adoptées pour explorer les facteurs influençant le quotient intellectuel (QI) et déterminer dans quelle mesure l'intelligence est le résultat de facteurs génétiques (innés) ou environnementaux (acquis).

Le code de notre projet est disponible à partir du lien Github suivant :

<https://github.com/ahmedmokeddem/dalas>.

# Cadrage du projet

---

Ce chapitre aborde le contexte de notre étude, en fournissant une vue d'ensemble complète du cadre dans lequel notre projet a été réalisé. Nous y présentons les raisons pour lesquelles cette recherche est pertinente, ainsi que l'environnement de travail et les données considérées et les méthodes de leur collecte.

## 1.1 Contexte

La question de savoir si l'intelligence est acquise ou innée est un sujet de débat de longue date qui traverse diverses disciplines telles que la psychologie, la génétique, l'éducation et les sciences sociales. Ce débat oppose deux perspectives principales : l'innéisme, qui soutient que l'intelligence est principalement déterminée par des facteurs génétiques, et l'empirisme, qui argue que l'intelligence est largement façonnée par l'environnement et l'expérience<sup>1</sup>.

Historiquement, les recherches sur l'intelligence ont souvent penché vers l'une ou l'autre de ces perspectives, avec des études mettant en avant l'importance des gènes et d'autres soulignant le rôle crucial des conditions environnementales telles que l'éducation, la nutrition et les interactions sociales.

### 1.1.1 Perspective innéiste

Les partisans de l'innéisme considèrent que l'intelligence est principalement déterminée par les gènes et l'hérédité. Selon cette vue, les individus naissent avec un certain "capital" intellectuel qui établit leurs capacités cognitives fondamentales<sup>2</sup>. Certains chercheurs estiment que l'intelligence est héritée à environ 80%. Cette perspective a été populaire au 19ème siècle, alimentée par des théories racistes et eugénistes visant à démontrer la supériorité intellectuelle de certains groupes<sup>3</sup>.

### 1.1.2 Perspective empiriste

À l'opposé, les tenants de l'empirisme soutiennent que l'intelligence est largement façonnée par l'environnement et l'expérience. Selon cette vue, le cerveau est très malléable, particulièrement durant l'enfance, et les stimulations environnementales jouent un rôle déterminant dans

1. [https://www.persee.fr/doc/raipr\\_0033-9075\\_1980\\_num\\_53\\_1\\_2035](https://www.persee.fr/doc/raipr_0033-9075_1980_num_53_1_2035)

2. [https://www.doctissimo.fr/html/sante/mag\\_2002/sem01/mag0524/dossier/sa\\_5529\\_inne\\_acquis.htm](https://www.doctissimo.fr/html/sante/mag_2002/sem01/mag0524/dossier/sa_5529_inne_acquis.htm)

3. <https://ludovicgadeau-psychotherapie.com/linneite-approche-philosophique/>

le développement cognitif<sup>4</sup>. Des études ont montré l'impact positif d'un environnement enrichi sur les capacités intellectuelles<sup>5</sup>.

## 1.2 Environnement de travail

Pour mener à bien notre projet, nous avons adopté un environnement de travail structuré et optimisé autour d'outils et de bibliothèques puissants et polyvalents. Voici un aperçu détaillé de notre configuration technique et des technologies utilisées :

### 1.2.1 Langage de programmation et éditeur

Notre code est intégralement écrit en **Python**, un langage de programmation largement utilisé dans le domaine de la data science pour sa simplicité, sa flexibilité et sa vaste collection de bibliothèques spécialisées. Pour le développement et l'exécution de notre code, nous avons utilisé **Jupyter Notebook** comme éditeur. Cet environnement interactif nous a permis de combiner du code, des visualisations et des annotations textuelles dans un même document, facilitant ainsi l'expérimentation, la documentation et le partage de nos analyses.

### 1.2.2 Collecte et structuration des données

Pour la collecte des données, nous avons utilisé plusieurs bibliothèques Python adaptées au web scraping :

- **Scrapy** : Une bibliothèque robuste et extensible pour le scraping de sites web à grande échelle. Scrapy nous a permis de définir des araignées (spiders) pour extraire systématiquement les données nécessaires.
- **BeautifulSoup** : Une bibliothèque pratique pour parser les documents HTML et XML. Nous l'avons utilisée pour extraire et naviguer facilement dans les données web.
- **Selenium** : Un outil puissant pour l'automatisation des navigateurs web. Selenium nous a été particulièrement utile pour interagir avec des pages web dynamiques et récupérer des données générées par JavaScript.

Une fois les données collectées, nous les avons structurées et manipulées en utilisant les bibliothèques suivantes :

- **NumPy** : Pour les opérations numériques et la manipulation de tableaux multidimensionnels.
- **Pandas** : Pour la manipulation et l'analyse des données sous forme de DataFrames, facilitant le nettoyage, la transformation et l'analyse des jeux de données.

4. <https://www.cairn.info/revue-devenir-2012-3-page-181.htm>

5. <https://www.telerama.fr/idees/etes-vous-heureux-en-meninges,n5218630.php>

### 1.2.3 Visualisation des données

Pour la visualisation des données, nous avons employé plusieurs bibliothèques spécialisées :

- **Plotly** : Pour des visualisations interactives et riches, permettant de créer des graphiques dynamiques et intuitifs.
- **Matplotlib** : Pour la génération de graphiques statiques, animés et interactifs en Python. Matplotlib est particulièrement utile pour les visualisations de base et personnalisées.
- **Seaborn** : Pour des visualisations statistiques attrayantes et informatives, basées sur Matplotlib mais avec une interface de haut niveau.

### 1.2.4 Analyse des données et modélisation

Pour l'analyse exploratoire des données (EDA) et le développement de modèles prédictifs, nous avons utilisé :

- **Scikit-learn (sklearn)** : Une bibliothèque de machine learning offrant un large éventail d'algorithmes pour la classification, la régression, le clustering et plus encore. Scikit-learn a été crucial pour l'EDA et pour la construction de nos modèles.
- **PyTorch (torch)** : Une bibliothèque de deep learning, offrant des outils pour le développement et l'entraînement de réseaux de neurones complexes. PyTorch nous a permis d'expérimenter avec des modèles plus sophistiqués et d'exploiter la puissance de l'apprentissage profond.

Pour les modèles de séries temporelles, nous avons utilisé :

- **Prophet** : Une bibliothèque développée par Facebook, spécialement conçue pour les prévisions de séries temporelles. Prophet est facile à utiliser et particulièrement efficace pour gérer des séries temporelles avec des composantes saisonnières et des tendances changeantes.

### 1.2.5 Interface utilisateur et déploiement

Pour la création d'une interface utilisateur interactive sous forme de tableau de bord, nous avons utilisé :

- **Streamlit** : Une bibliothèque Python qui simplifie la création d'applications web interactives. Streamlit nous a permis de développer un tableau de bord interactif pour la visualisation et l'exploration des résultats de nos analyses et modèles de manière intuitive et accessible.

Pour le déploiement de notre application, nous avons opté pour :

- Docker : Une plateforme de conteneurisation qui nous a permis d'emballer notre application et toutes ses dépendances dans un conteneur unique. Docker a facilité le déploiement de notre tableau de bord et de nos modèles dans différents environnements, assurant une portabilité et une reproductibilité optimales.

## 1.3 Acquisition des données

Pour mener à bien notre projet, il était essentiel de collecter des données pertinentes et fiables. Nous avons adopté une approche exhaustive et méthodique pour rassembler les données nécessaires à notre analyse. Voici une description détaillée des attributs utilisés, leurs sources (avec le lien pour y accéder) ainsi que leurs descriptions.

Il est à noter que certaines données ont été scrapé à partir du site web source et d'autres ont été récupérées en format csv. Aussi, nous avions prévu de scrapper les tweets sur l'éducation nationale dans chacun des pays pour les analyser mais nous n'avons pas pu ni en scrapping ni en utilisant l'API proposée par X (anciennement Twitter)

Attribut	Source	Scrapé	Description
Country	World Population Review	Non	Indique le pays
Notes_Musees	Google Maps	Oui	Les notes des principaux musées de chaque pays
Nbvotes_Musees_Clean	Google Maps	Oui	Nombre de personnes ayant noté les musées de chacun des pays
Nb_Prixnobel	Wikipedia	Oui	Indique le nombre de prix nobels obtenus par chaque pays
Annee_Souverainete	Wikipedia	Oui	L'année de l'indépendance de chaque pays
Nb_Univtop500	Top Universities	Oui	Indique le nombre d'universités dans le top 500 du classement des universités
Mean_Rank_Univ	Top Universities	Oui	Classement moyen des universités de chaque pays
Political_Regime	our world in data	Non	le régime politique de chaque pays

Immigrationbycountry_Immigrants	world population review	Non	Taux d'immigrés parmi la population de chaque pays
Immigrationbycountry_Emigrants	world population review	Non	Taux d'émigrés parmi la population de chaque pays
Primary	OCDE	Non	Nombre d'heures enseignées au primaire
Lower_Secondary	OCDE	Non	Nombre d'heures enseignées au collège
Break_1	OCDE	Non	Nombre de jours de la première vacance
Break_2	OCDE	Non	Nombre de jours de la deuxième vacance
Break_3	OCDE	Non	Nombre de jours de la troisième vacance
Break_4	OCDE	Non	Nombre de jours de la quatrième vacance
Break_5	OCDE	Non	Nombre de jours de la cinquième vacance
End_Of_The_School_Year_Break	OCDE	Non	Nombre de jours des vacances d'été
Literacy_Rate_2021	The global economy	Non	Indique le taux d'alphabétisme par pays
Global_rank_Literacy_rate	The global economy	Non	Indique le classement de chaque pays par rapport à son taux d'alphabétisme
Continent	Clearias	Oui	Indique le continent auquel appartient chacun des pays
Area	Wikipedia	Oui	Indique la superficie de chacun des pays
Nb_graduates	OCDE	Non	Indique le nombre de diplômés de chaque pays
Population	worldbank	Oui	Indique la population de chacun des pays

Education_Spending_2021	The global economy	Non	Indique le pourcentage du budget de l'éducation nationale par pays
Gdp	worldbank	Oui	Indique le PIB (Produit intérieur brut)
Averageiqbycountry_Iqlynnbecker2019	World Population Review	Non	Indique le QI moyen de chaque pays
Averageiqbycountry_Sourcelynnbecker2019	World Population Review	Non	Indique si les données sur le pays permettant le calcul du QI sont disponible (sinon le QI est estimé)
Averageiq_Ici2017Score	World Population Review	Non	Indique le score ICI (Intelligence Capital Index) de chaque pays
Averageiqpisa2022Meanscoremathematics	World Population Review	Non	Note moyenne du classement Pisa en mathématiques
Averageiqpisa2022Meanscorereading	World Population Review	Non	Note moyenne du classement Pisa en lecture
Averageiqpisa2022Meanscorescience	World Population Review	Non	Note moyenne du classement Pisa en science
Gdp_Percapita	Attribut calculé	/	Indique le PIB par habitant pour chaque pays
Indice_Evolution	Attribut calculé	/	Indique l'indice d'évolution qui est un rapport entre le PIB et le nombre d'années depuis l'indépendance
Nb_Foreign_Students	OCDE	Non	Pourcentage d'étudiants étrangers parmi les étudiants nationaux de chaque pays

Indice_culturel	Attribut calculé	Non	Indice de notation calculé en multipliant le nombre de notes sur les musées par la note moyenne des musées
-----------------	------------------	-----	--

Notons que l'année considérée pour chaque attribut du jeu de données est 2019 (ou l'année qui s'en rapproche le plus), car celle-ci correspond à l'année à laquelle le test de QI a été effectué.

# Analyse de données

---

Ce chapitre est dédié à l'analyse préliminaire et exploratoire des données. Ainsi, nous affichons des graphiques tels que la distribution des données brutes, le pourcentage de valeurs nulles des attributs etc, puis nous décrivons le processus de nettoyage et de normalisation des données, et enfin, analysons et interprétons les données à travers quatre axes : « géoéconomique », « historique », « socio-économique », « culturel et éducatif ». Nous terminons par une analyse en composantes principales (ACP) afin de voir comment sont liés tous les attributs au Quotient Intellectuel ; le QI étant au cœur de notre étude.

## 2.1 Analyse préliminaire

Après avoir récupéré toutes les données susmentionnées, celles-ci passent par une étape de prétraitement essentielle. Ce titre aborde cette phase cruciale qui consiste à analyser la distribution des valeurs des variables, identifier et traiter les valeurs manquantes, détecter les anomalies, et effectuer les transformations nécessaires pour préparer les données pour les analyses ultérieures.

### 2.1.1 Distribution des variables

A titre indicatif, nous considérons ici certaines des variables numériques.

Nous pouvons avoir les caractéristiques statistiques des variables en utilisant les méthodes de Pandas. Ainsi, pour certaines variables, nous pouvons remarquer que beaucoup de valeurs sont manquantes. C'est le cas de la variable Primary pour laquelle uniquement 38 valeurs sont renseignées sur un total de 197. Ce tableau nous indique également la valeur moyenne pour chacune des variables. Ainsi, le nombre de prix nobels moyen par pays est 6, en moyenne 2 universités sont dans le top500 des universités pour chacun des pays. Aussi, nous pouvons aussi analyser les valeurs minimales. Par exemple, le pays dont l'année de souveraineté est le plus ancien a pour année de souveraineté 843. En ce qui concerne les valeurs maximales, l'année de souveraineté le plus récent est en 2011. En termes de prix Nobels, les États Unis est celui ayant le plus de prix avec 411 prix.

Nb_Prixnobel	Annee_Souverainete	Nb_Univtop500	Immigrationbycountry_Immigrants	Immigrationbycountry_Emigrants	Primary
197.000000	164.000000	197.000000	1.920000e+02	1.920000e+02	38.000000
6.081218	1887.554878	2.522843	1.452172e+06	1.388475e+06	4400.263158
32.515044	193.145174	8.130049	4.263757e+06	2.259164e+06	1375.457087
0.000000	843.000000	0.000000	2.139000e+03	1.908000e+03	1890.000000
0.000000	1884.750000	0.000000	5.737900e+04	1.520702e+05	3452.000000
0.000000	1956.000000	0.000000	2.889990e+05	6.938305e+05	4502.000000
1.000000	1968.000000	1.000000	1.170911e+06	1.599130e+06	5379.250000
411.000000	2011.000000	82.000000	5.063284e+07	1.786949e+07	7000.000000

FIGURE 2.1 – Caractéristiques statistiques des variables

Pour analyser plus en détails ces variables, nous considérons, à titre d'exemple, certains diagrammes à moustaches.

Celui de la variable indice évolution est représenté par la figure qui suit. Nous pouvons voir que l'échelle des valeurs est entre 0 et 370, les valeurs au dessus sont considérées comme des valeurs aberrantes. C'est le cas par exemple de pays comme le Singapour pour lequel l'année de souveraineté est assez récente et dont le GDP est élevé. Ceci indique concrètement que ce pays, bien qu'il ait été colonisé, a pu se relever et reconstruire son économie. Ainsi, les valeurs aberrantes sont liées aux pays dont l'année d'indépendance est récent ou alors leurs PIB est très élevé.

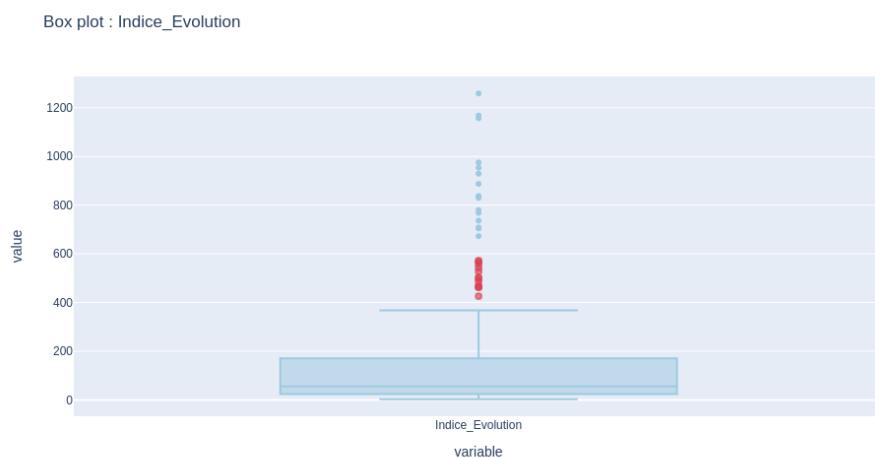


FIGURE 2.2 – Distribution de la variable indice\_evolution

En ce qui concerne la variable PIB, nous pouvons remarquer que les valeurs se situent majoritairement entre 9 millions et 600 milliards. Néanmoins, nous remarquons aussi des valeurs aberrantes très élevées. C'est le cas des états-unis et de la Chine dont le PIB s'élève à, respectivement, 26k et 21k milliards de dollar. Ceci indique une grande variance entre les valeurs de cette variable.

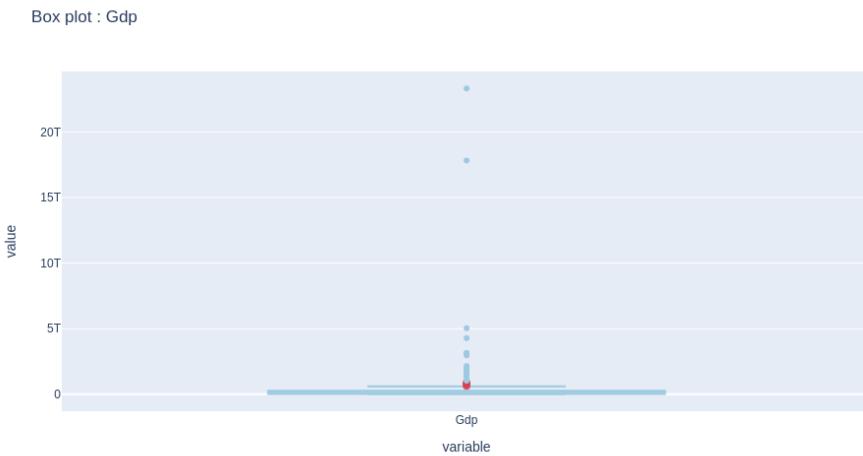


FIGURE 2.3 – Distribution de la variable GDP

La variable population est aussi sur une échelle de valeurs élevées. Nous remarquons que la valeur médiane est 9 millions. La Chine compte 1.5 Milliards d'habitants ce qui fait d'elle le pays le plus peuplé du monde. Nous pouvons aussi constater que 96% des pays ont une population inférieure à 71 millions d'habitants.

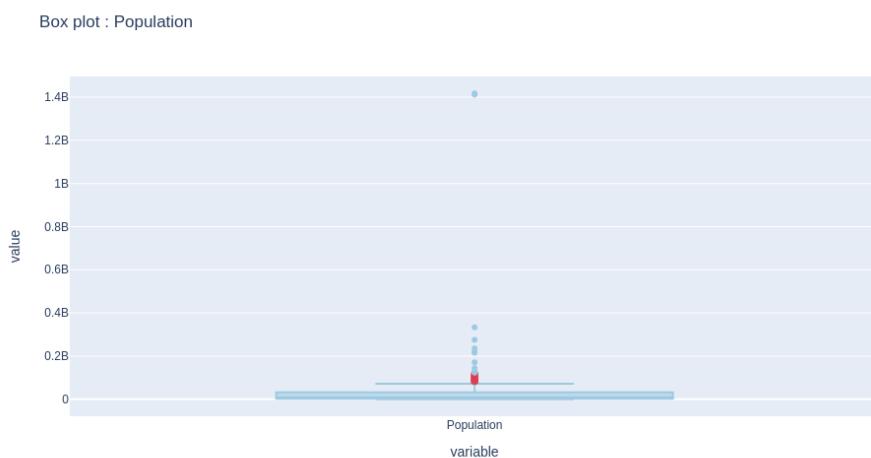


FIGURE 2.4 – Distribution de la variable Population

Le taux d'alphabétisme dans chacun des pays est situé entre 0 et 100%. Nous remarquons que 96% des pays ont un taux d'alphabétisme entre 89 et 100%. Néanmoins, il existe des pays dont le taux d'alphabétisme est faible jusqu'à atteindre 37%.

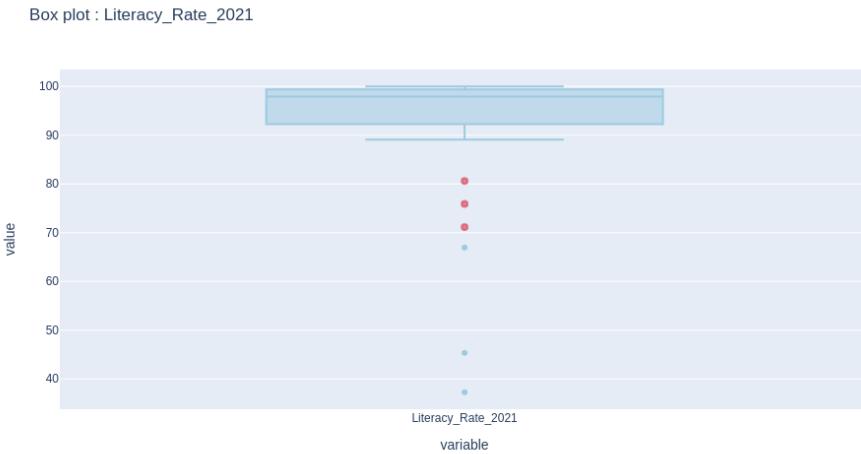


FIGURE 2.5 – Distribution de la variable taux d’alphabétisme

Nous remarquons que 96% des pays investissent moins de 8% de leurs PIB dans l’éducation nationale. Avec des valeurs aberrantes, c’est notamment le cas de la Chine.

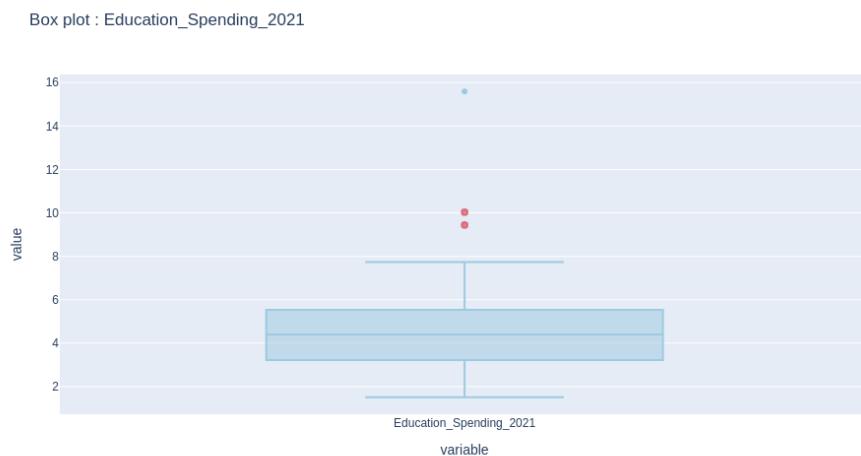


FIGURE 2.6 – Distribution de la variable Budget éducation nationale

Quant à l’année de souveraineté, nous remarquons que la majorité des pays ont obtenu leur souveraineté depuis l’an 1768. Ainsi, l’indépendance la plus récente que nous avons enregistré date de 2011.

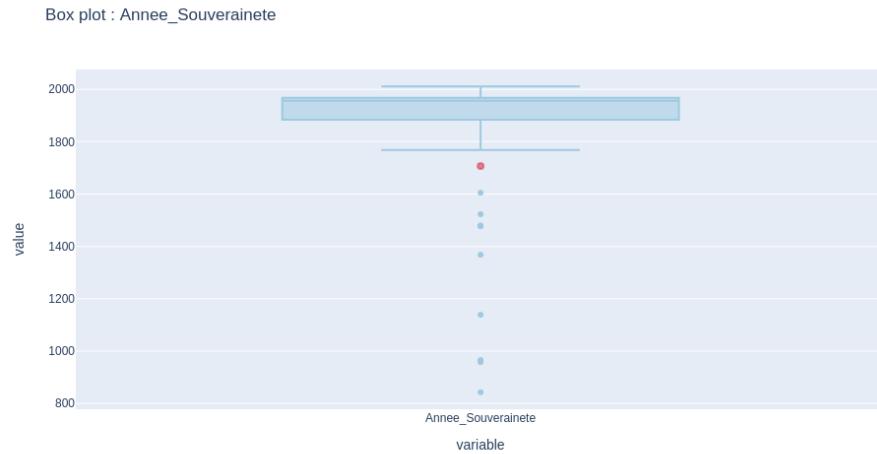


FIGURE 2.7 – Distribution de la variable Année de souveraineté

Concernant les étudiants étrangers, nous remarquons qu'en moyenne 10% des étudiants de chaque pays est étranger. Le Luxembourg se distingue par un pourcentage approchant les 50% soit la moitié des étudiants du Luxembourg sont étrangers.

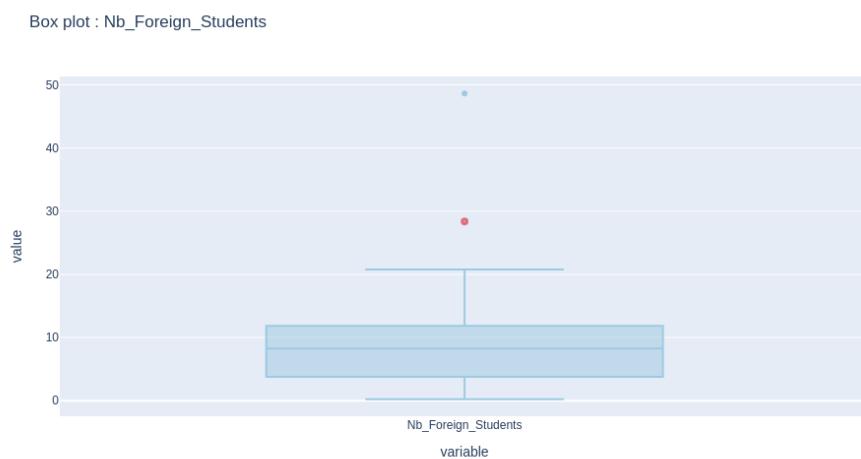


FIGURE 2.8 – Distribution de la variable Nombre d'étudiants étrangers

Comme on peut les remarquer, les échelles diffèrent d'une variable à une autre. En effet, les valeurs prises par la variable population sont très élevées par rapport aux valeurs prises par la variable pourcentage étudiants étrangers. De ce fait, dans la suite de nos analyses nous les standardisons en retranchant la moyenne et divisant par l'écart type. Ainsi, la boîte à moustache qui suit illustre les diagrammes correspondant à chacune des variables.

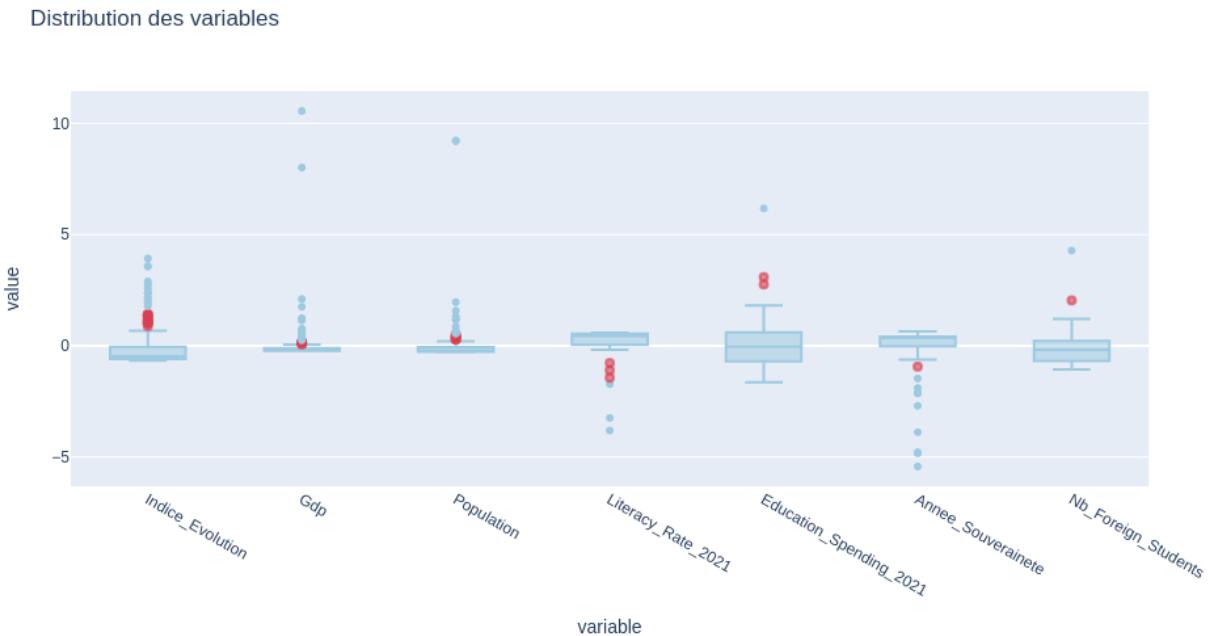


FIGURE 2.9 – Distribution de certaines variables normalisées

Pour analyser les relations entre les différentes variables, nous avons tracé une matrice de corrélation linéaire. Il est évident que la diagonale présente des corrélations parfaite puisqu'une variable est corrélée avec elle-même. Nous pouvons relever certaines corrélations élevées entre les variables pourcentage immigrés et nombre de prix Nobel avec une corrélation de 0.9, entre les variables nombre de diplômés et PIB, ou encore entre le budget de l'éducation nationale et le PIB. Cela signifie que ces variables évoluent dans le même sens, plus l'une croît, l'autre croît également. D'autres variables sont négativement corrélées, c'est le cas par exemple entre les variables nombre de diplômés et classement suivant le taux d'alphabétisme. Cela suggère que plus le nombre de diplômés est élevé, le pays obtient les premiers rangs du fait d'un taux d'alphabétisme élevé.

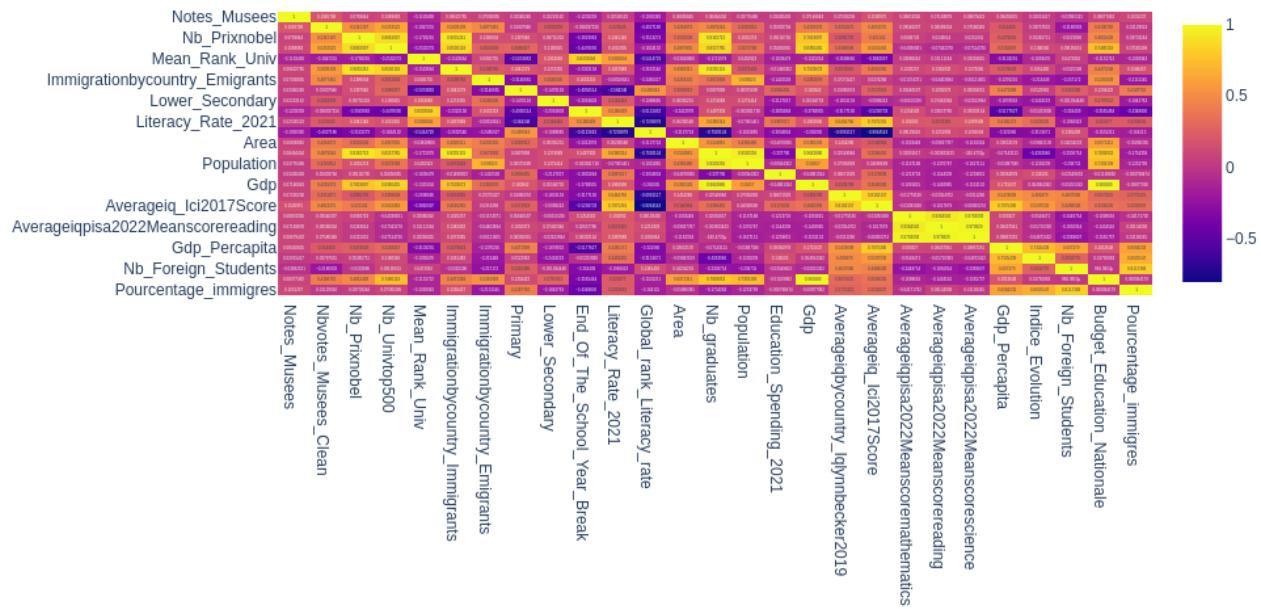


FIGURE 2.10 – Corrélation linéaire entre les variables

### 2.1.2 Valeurs renseignées

Pour vérifier si une variable est pertinente pour notre étude, celle-ci doit avoir suffisamment de valeurs pour pouvoir l'analyser. Ainsi, la figure 2.11 indique le pourcentage des valeurs manquantes pour chacune des variables considérées dans notre étude.

Nous remarquons que les variables indiquant les vacances scolaires pour les élèves présentent des pourcentages de valeurs manquantes très élevés, ainsi, nous les excluons de notre étude car les seules valeurs renseignées ne nous permettent pas d'établir leurs effets sur la valeur du QI dans le cadre de notre étude. Ceci est aussi le cas pour les variables indiquant le nombre des étudiants étrangers, classement moyen des universités et nombre de diplômés. Pour les autres variables des techniques de remplacement sont utilisées pour renseigner les valeurs manquantes, c'est le cas, par exemple, de la variable GDP pour laquelle lorsque la dernière valeur n'est pas disponible, nous considérons la dernière valeur disponible.

Country	0.000000
Notes_Musees	0.000000
Nbvotes_Musees_Clean	0.000000
Nb_Prixnobel	0.000000
Annee_Souverainete	16.751269
Nb_Untivtop500	0.000000
Mean_Rank_Univ	70.558376
Political_Regime	14.720812
Immigrationbycountry_Immigrants	2.538071
Immigrationbycountry_Emigrants	2.538071
Primary	80.710660
Lower_Secondary	80.710660
Break_1	83.248731
Break_2	83.248731
Break_3	86.802030
Break_4	92.893401
Break_5	98.984772
End_Of_The_School_Year_Break	83.248731
Literacy_Rate_2021	80.710660
Global_rank_Literacy_rate	80.710660
Continent	0.000000
Area	3.553299
Nb_graduates	85.279188
Population	2.538071
Education_Spending_2021	29.949239
Gdp	0.000000
Averageiqbycountry_Iqlynnbecker2019	0.000000
Averageiqbycountry_Sourcelynnbecker2019	0.000000
Averageiq_Ici2017Grade	35.532995
Averageiq_Ici2017Score	35.532995
Averageiqpisa2022Meanscoremathematics	59.898477
Averageiqpisa2022Meanscorereading	59.898477
Averageiqpisa2022Meanscorescience	59.898477
Gdp_Per capita	3.045685
Indice_Evolution	17.766497
Nb_Foreign_Students	81.725888

FIGURE 2.11 – Pourcentage de valeurs manquantes par variable

## 2.2 Analyse Exploratoire des Données (EDA)

Nous présentons dans cette partie les axes d’analyses choisis pour l’étude de notre problématique intitulée « L’intelligence innée ou acquise ? ». Pour chaque axe d’analyse présenté, nous affichons les graphiques correspondants, ainsi que l’interprétation que nous pouvons en déduire.

### 2.2.1 Axe géoéconomique

Cet axe à pour but d’explorer la production de richesse sur une surface géographique donnée. Nous avons choisis d’étudier cela sur deux granularités différentes ; la première par rapport aux continent, et la seconde, plus fine, se fait par rapport aux pays.

PIB moyen par continent

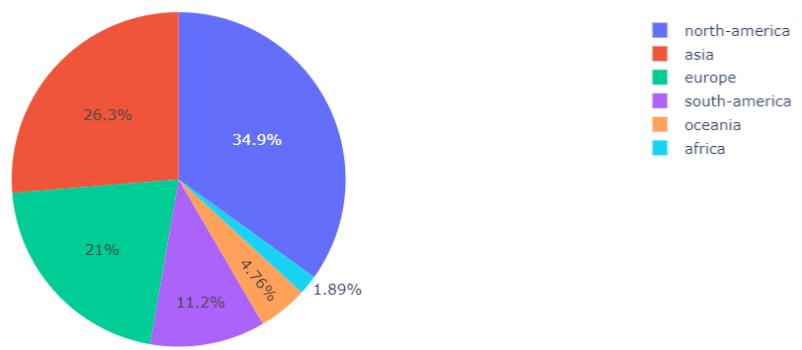


FIGURE 2.12 – PIB moyen de par continent

Le Produit Intérieur Brut (PIB) moyen pour chaque continent montre que les 80% de l’activité économique est occupée par les pays d’Amérique du nord, d’Asie et d’Europe.

Sauf qu’on peut remarquer une variance entre les production des pays d’un même continent, c’est pour cela que nous introduisons une analyse plus fine qui consiste à afficher dans l’ordre le pourcentage de richesse des pays produisant plus de 2 fois le PIB moyen global.

PIB moyen par pays

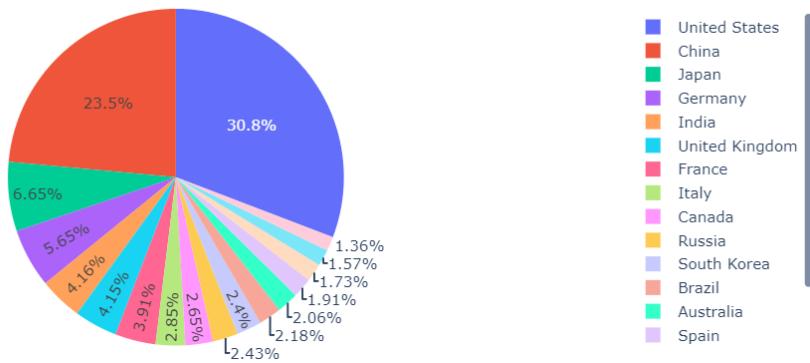


FIGURE 2.13 – PIB moyen de par pays

Nous retrouvons dans le top 5 les USA suivit de la Chine, a eux deux cumules plus de 50% de l’économie de la liste des pays retenus. Puis viens successivement le Japon, l’Allemagne et l’Inde.

Cette disparité entre les économies peut avoir une relation avec le passé de chaque pays, c’est ce que nous étudions dans le titre suivant.

### 2.2.2 Axe historique

L’objectif à travers cet axe est d’intégrer une dimension temporelle à notre analyse exploratoire des données. La dimension temporelle est traduite par le passé colonial du pays et/ou son évolution économique à travers les années.

## — Evolution du Produit Intérieur Brut (PIB)

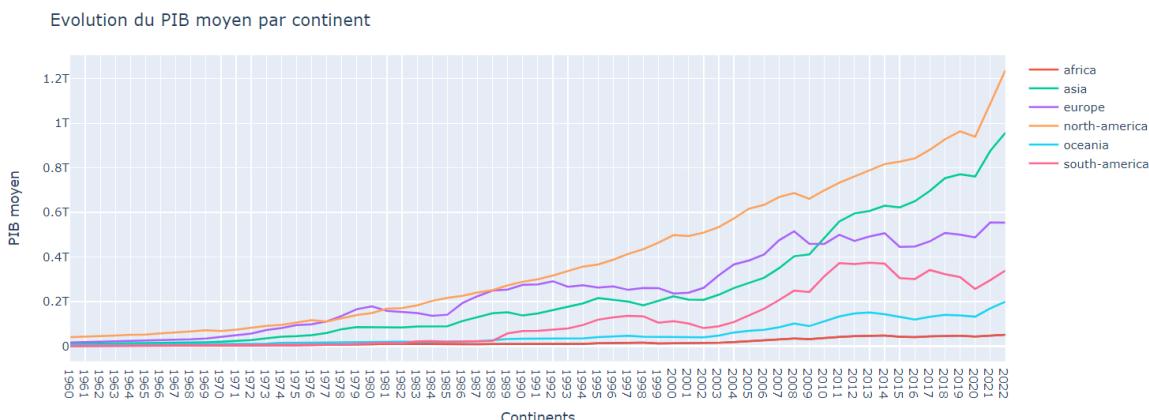


FIGURE 2.14 – Evolution du PIB

Avant 1988, l'Europe et l'Amérique du Nord se disputaient la première position en termes de PIB, avec l'Amérique du Nord finissant par prendre l'avantage. Durant cette période, l'Amérique du Sud a montré une progression notable, passant de la dernière position, à égalité avec l'Afrique, pour rejoindre et dépasser l'Océanie. L'Asie, pour sa part, a maintenu une position intermédiaire, occupant la troisième place. À partir de 2010, l'Asie a connu une ascension remarquable, se hissant à la deuxième position, juste derrière l'Amérique du Nord, et devançant ainsi l'Europe.

Cette analyse nous donne une idée sur la richesse économique moyenne distribuée par continent. Cependant, notons que cette richesse est relative à la surface géographique des pays qui appartiennent aux continents, au nombre de leurs habitants mais aussi, et surtout, à leur passé coloniale. En effet, ces trois paramètres influencent particulièrement l'activité économique d'un pays, car si par exemple un pays n'a repris sa souveraineté qu'en 1975 alors comparer sa richesse avec celle du pays colonisateur n'est pas réellement indicateur du potentiel économique des deux pays. C'est pour cette raison que nous introduisons l'**indice d'évolution** =  $\frac{PIBparhabitant}{annécourante - annéessouveraineté}$ , plus cet indice est élevé et plus on peut dire qu'un pays évolue rapidement, nous analyserons cela dans les prochains graphiques.

## — Relation entre l'indice d'évolution et l'année d'indépendance

Nous étudions à travers le graphique 2.15 la relation entre l'indice d'évolution et l'année d'indépendance des pays de chaque continent.

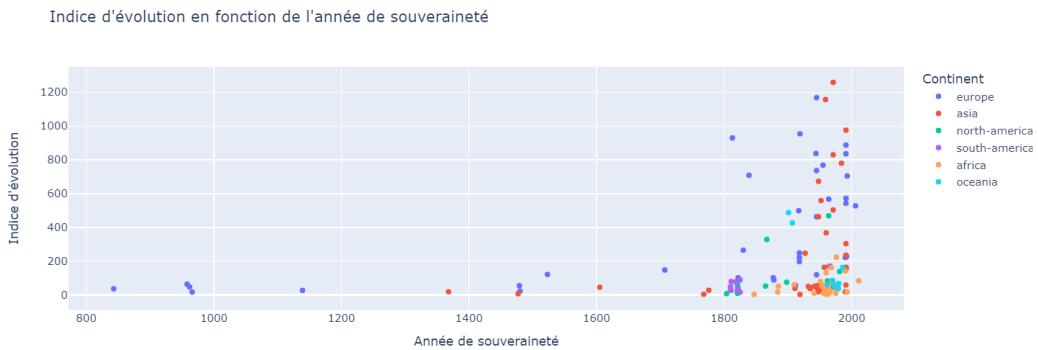


FIGURE 2.15 – Indice d'évolution et l'année d'indépendance par continent

Nous observons qu'avant 1368, seuls des pays européens avaient la totale possession de leurs patrimoines. Cependant, leur indice d'évolution est relativement bas, ce qui signifie que leur PIB par habitant est certes probablement élevé mais pas assez pour montrer une réelle évolution du pays.

Ce n'est qu'à partir des années 1800 que nous remarquons un indice d'évolution plus ou moins important. Trois catégories sont à distinguer, la première est celle de certains pays d'Asie, qui aux côtés des pays d'Europe montrent la meilleure évolution. La seconde catégorie concerne les pays d'Océanie et d'Amérique du nord, ces derniers sont à une échelle d'évolution moyenne. Enfin, nous retrouvons majoritairement les pays d'Afrique, d'Amérique du sud et des pays d'Asie, qui bien que leur année d'indépendance soit récente, leur PIB par habitant n'est pas assez élevé pour impacter significativement leur indice d'évolution.

### — Indice d'évolution par continent

Nous agrégeons les résultats précédents par le graphique 2.16.

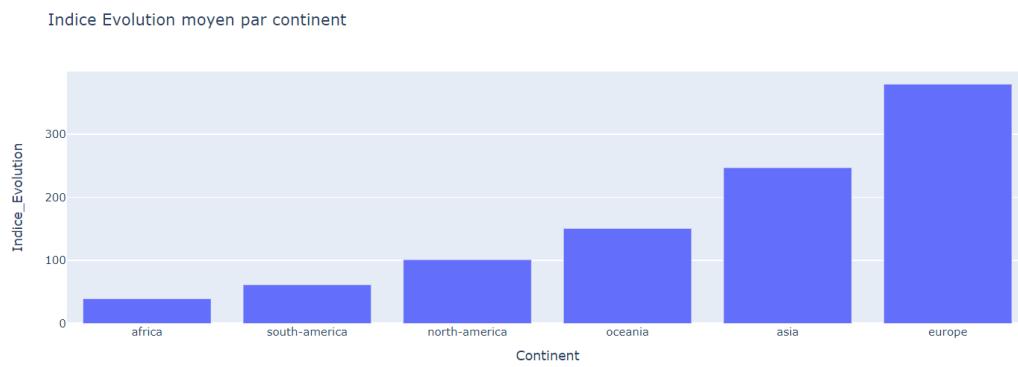


FIGURE 2.16 – Indice évolution moyen par continent

Les continents qui comportent en moyenne les pays ayant connu la meilleure évolution sont successivement : l'Europe, l'Asie, l'Océanie, l'Amérique du nord, l'Amérique du sud, l'Afrique.

### 2.2.3 Axe socio-économique

Notre étude à travers cet axe vise à lier la dimension sociale à la dimension économique des pays. Nous explorons pour cela les corrélations possibles entre des paramètres tels que le régime politique, le PIB, le budget alloué à l'éducation nationale, le nombre d'immigrés/émigrants.

#### — Relation entre le budget d'éducation nationale et le PIB

L'idée à travers cette analyse est de voir si l'éducation est au cœur des préoccupations d'un pays donné. Pour ce premier graphique 2.17 nous remarquons la distinction flagrante de la Chine, avec un PIB et un budget d'éducation nationale très élevés.

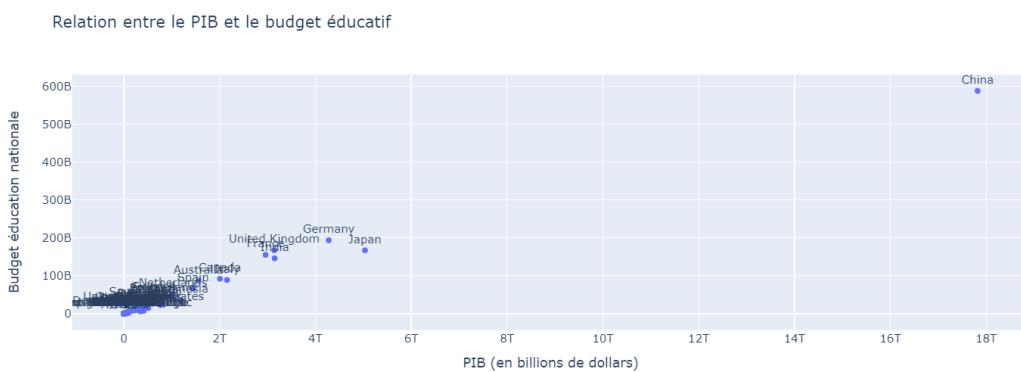


FIGURE 2.17 – Budget alloué à l'éducation nationale et PIB (1)

Zoomer sur le graphique fait apparaître les pays tel que le Japon, l'Allemagne ou encore la France, le Royaume-Uni et l'Inde.

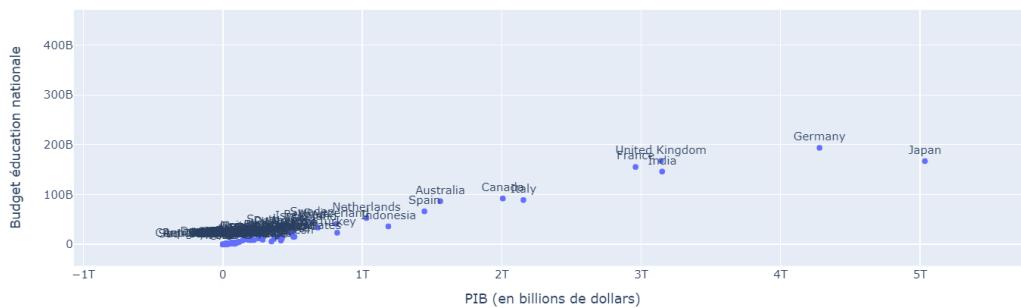


FIGURE 2.18 – Budget alloué à l'éducation nationale et PIB (2)

Pour approximativement une même valeur de PIB, on peut voir que la Suède investit plus de budget pour l'éducation nationale que la Pologne et la Belgique. Un paramètre supplémentaire qu'il serait pertinent de prendre en compte serait le taux d'étudiants de chaque pays par rapport à sa population.

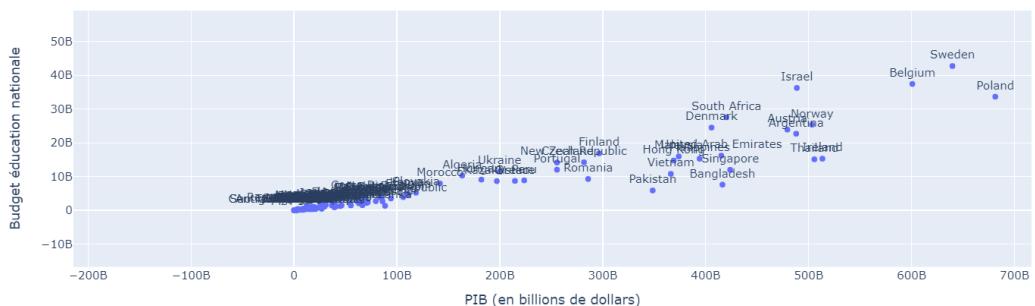


FIGURE 2.19 – Budget alloué à l'éducation nationale et PIB (3)

Nous pouvons globalement conclure que les deux facteurs, budget d'éducation nationale et PIB, évoluent de manière proportionnelle.

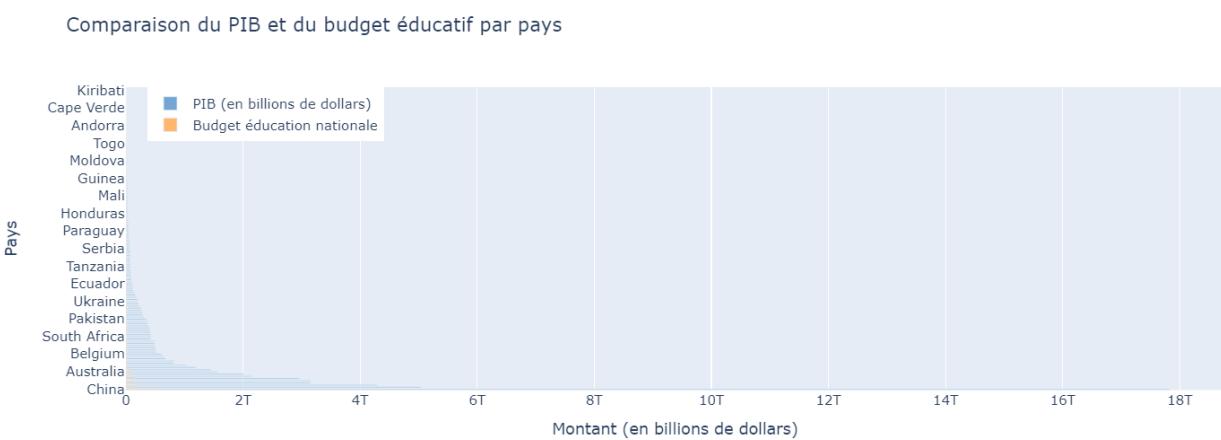


FIGURE 2.20 – Budget alloué à l'éducation nationale et PIB (4)

Les figures 2.20 et 2.21 montrent le budget (en billions de dollars) que chaque pays alloue à son éducation nationale et le compare son PIB. Ici aussi, on peut observer que la Suède investit plus de budget pour l'éducation nationale que la Pologne et la Belgique. L'avantage avec ce type de graphique, contrairement au précédent qui montrait principalement la proportionnalité des deux paramètres, est qu'on peut dé-sélectionner (graphique interactif) un des deux paramètres pour ne comparer par exemple que les budgets d'éducation nationale entre les pays.

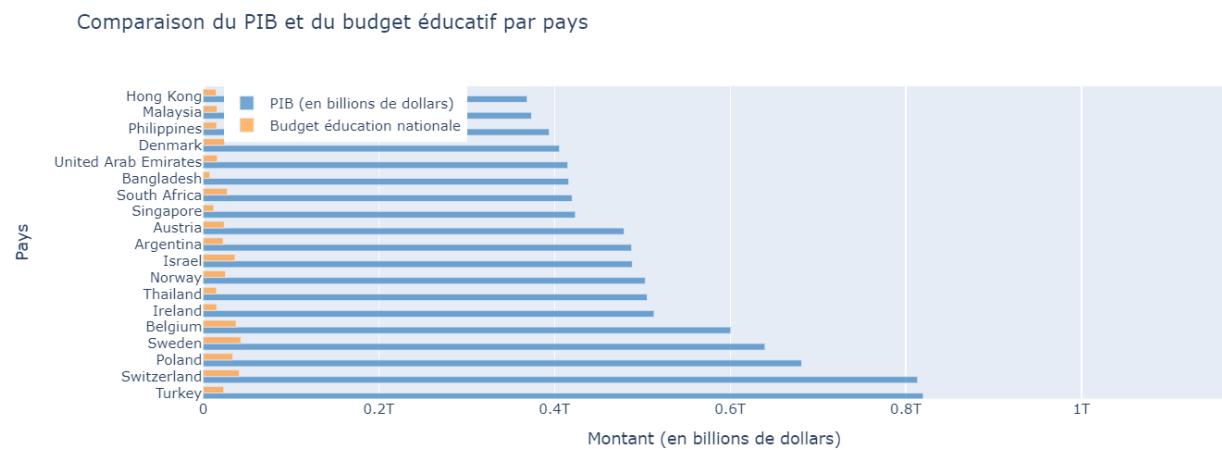


FIGURE 2.21 – Budget alloué à l'éducation nationale et PIB (5)

Notons que la première figure 2.20 n'est à priori pas très clair du fait que le PIB de la Chine soit important. Pour une meilleure visibilité, nous aurions pu comparer que les pourcentages du budget d'éducation nationale (par rapport au PIB) et comparer les pays entre eux ; sauf qu'avec cette approche nous aurions été incapable de comparer deux pays avec un même PIB mais une différence flagrante du budget alloué à l'éducation nationale. De plus, nous n'aurions eu que des valeurs relatives au PIB ce qui peut fausser les interprétations car les pays avec un PIB très important se retrouveraient forcément pénalisés, alors qu'il se peut que le budget alloué à l'éducation national soit tout aussi important et suffisant à une bonne prise en charge des étudiants.

### — Impact du régime politique sur le budget d'éducation nationale

L'analyse précédente (Figure 2.19) a montré la proportionnalité entre le budget d'éducation nationale et le PIB. Un facteur supplémentaire qu'il serait pertinent de considérer est le régime politique d'un pays ; car en fonction d'un certain régime politique, il se peut qu'un pays n'investisse pas grandement des ressources de l'état, mais plutôt compte sur le secteur privé pour développer l'éducation de ses citoyens.

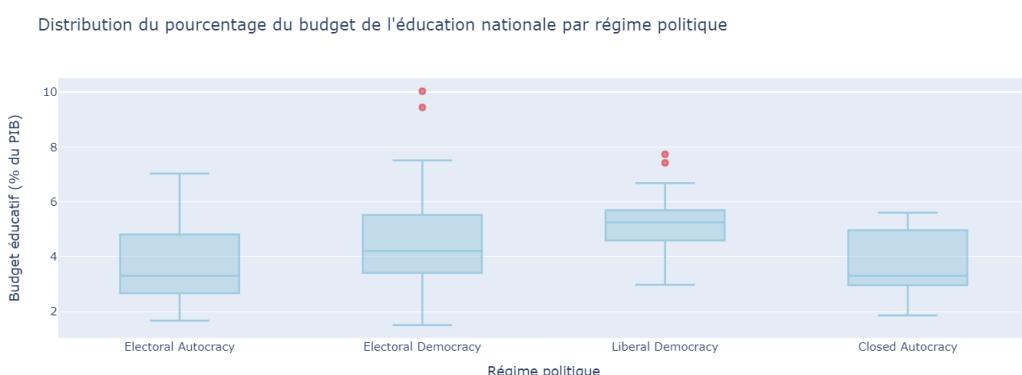


FIGURE 2.22 – Régime politique et Budget d'éducation nationale

Ainsi, à partir de la de la figure 2.22 et en considérant les valeurs médianes et quartiles des

budgets alloués à l'éducation nationale (en % par rapport au PIB de chaque pays), nous pouvons déduire que les pays dont le régime politique suit une « Démocratie libérale » sont de plus faible variance quant à la priorisation de l'éducation nationale. Puis on retrouve la « Démocratie électorale » suivit des deux type d'Autocraties.

### **Remarque :**

Les différences entre ces régimes politiques sont donnés par le site<sup>1</sup> comme suit :

- Autocraties fermées : Dans ce régime, les citoyens n'ont pas le droit de choisir le chef de l'exécutif du gouvernement ni la législature par le biais d'élections multipartites.
- Autocraties électorales : Dans ce régime, les citoyens ont le droit de choisir le chef de l'exécutif et la législature par le biais d'élections multipartites, mais ils manquent de certaines libertés, comme la liberté d'association ou d'expression, qui rendent les élections significatives, libres et équitables.
- Démocraties électorales : Dans ce régime, les citoyens ont le droit de participer à des élections multipartites significatives, libres et équitables.
- Démocraties libérales : Dans ce régime, les citoyens bénéficient de droits individuels et des minorités supplémentaires, sont égaux devant la loi, et les actions de l'exécutif sont contraintes par le législatif et les tribunaux.

### **— Analyse de données multivariée**

Dans cette partie, nous analysons la relation entre le PIB, le Nombre d'émigrants/immigrant et Quotient Intellectuel (QI) par pays et par continent.

Grâce aux deux graphiques 2.23 et 2.24 suivants nous pouvons situer un pays par rapport à un autre sur deux dimensions. De plus, nous pouvons observer la dispersion des pays au sein d'un même continent par rapport à différents facteurs étudiés pair à pair.



FIGURE 2.23 – Relation entre le QI PIB et le nombre d'immigrants/émigrants par pays

La Chine et les USA se distingue particulièrement par rapport aux autres pays lorsqu'il s'agit d'évaluer à la fois le QI et le PIB du pays, sauf qu'en observant le nombre d'émigrants

1. <https://ourworldindata.org/grapher/political-regime>

et immigrés de ces deux pays, la différence est significative. Cela peut nous faire penser que la richesse intellectuelle des USA n'est peut être pas liée à son système éducatif ou à la génétique de sa population mais plutôt au phénomène dit de la «fuite des cerveaux<sup>2</sup>», contrairement à un pays comme la Chine.

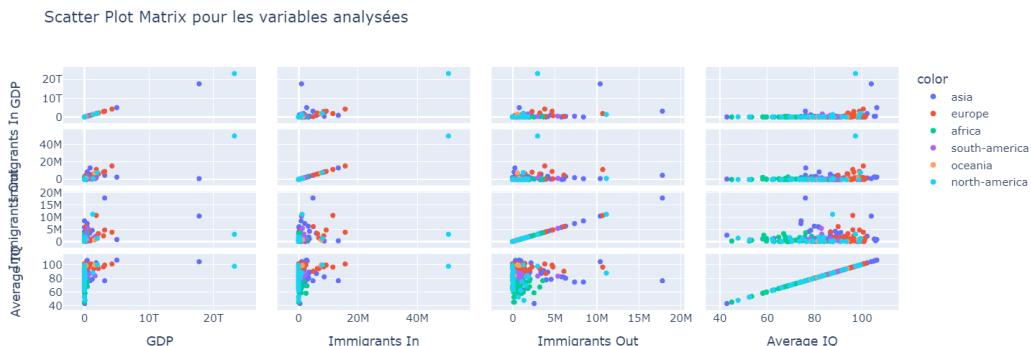


FIGURE 2.24 – Relation entre le QI PIB et le nombre d'immigrants/émigrants par continent

Nous remarquons une certaine homogénéité entre les pays d'un même continent puisqu'ils sont plus ou moins regroupés et répartis similairement sur les différents axes d'analyse de la figure 2.24.

## 2.2.4 Axe culturel et éducatif

A travers cet axe, nous voulons voir si un pays peut être concerné par le phénomène décrit précédemment dit de la «fuite des cerveaux», tout comme nous cherchons à analyser l'impact de facteurs comme l'indice culturel d'un pays sur l'éducation de ses citoyens.

### — Proportion d'immigrés et d'étudiants étrangers

Nous avons soulevé un point important lors de l'analyse précédente (voir partie 2.2.3), ce point concerne la migration sélective de personnes hautement qualifiées vers des pays où les conditions de vies semblent plus favorables. C'est pour cette raison que nous avons voulu calculer le pourcentage d'étudiants étrangers (par rapport aux étudiants locaux) ainsi que calculer le pourcentage d'immigrés (par rapport à la population totale) afin de voir si un pays peut être potentiellement concerné par le *brain gain* ou « attraction des compétences ».

2. une forme de migration sélective internationale de personnes hautement qualifiées

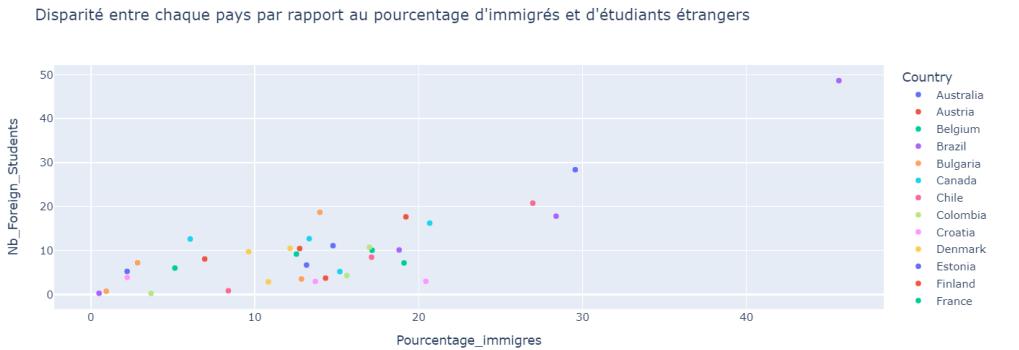


FIGURE 2.25 – Relation entre le nombre d'étudiants étrangers et le nombre d'immigrés par pays

La première analyse 2.25 montre une proportionnalité entre le pourcentage d'immigré et celui du nombre d'étudiants étrangers. Ceci nous laisse penser que les étudiants étrangers restent après leurs études ce qui peut participer notamment à augmenter le quotient intellectuel (QI) moyen de la population, et que ce dernier n'est peut être pas lié à forcément au système éducatif ou à la génétique de la population d'un pays mais plutôt aux conditions de vies en général du pays.

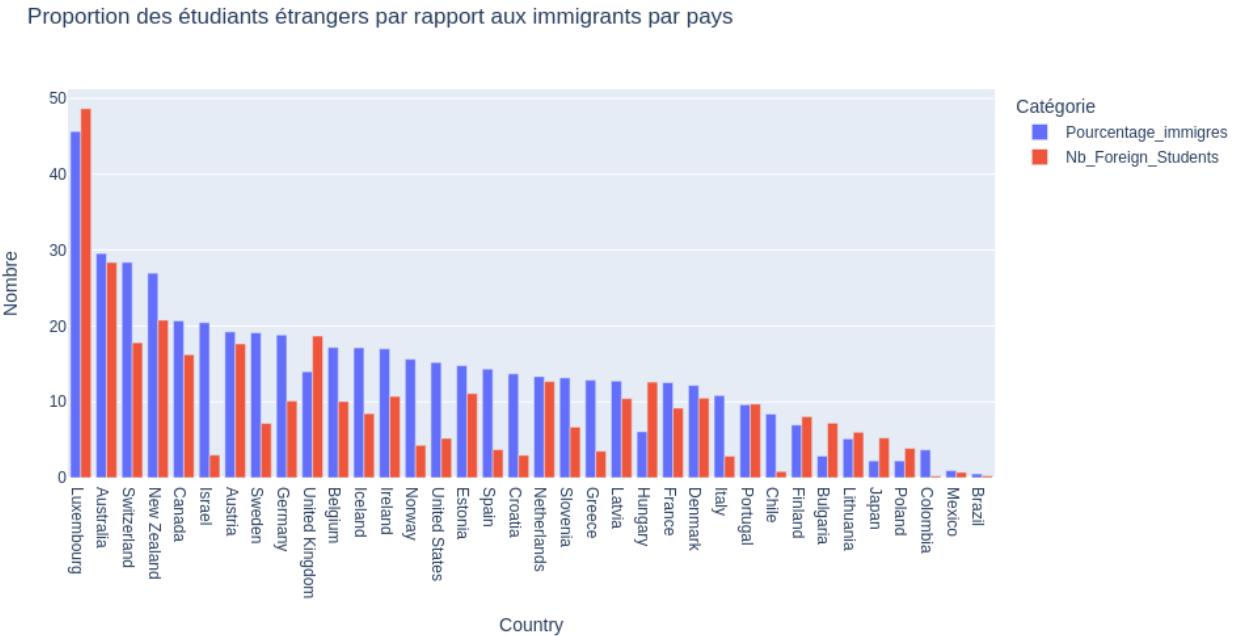


FIGURE 2.26 – Proportion des étudiants étrangers par rapport aux immigrés par pays

La figure 2.26 permet de mieux distinguer et comparer les différents taux étudiés entre les pays. Nous pouvons observer par exemple que le Luxembourg comprend presque 50% d'immigrés parmi sa population totale et le ratio de ses étudiants étrangers est encore plus élevé, cela peut indiquer que le pays est plus ouvert (et/ou attire davantage) aux étrangers qu'un pays comme le Chili, pour qui les proportions sont beaucoup moins importantes.

### — Relation entre l'indice culturel et le taux d'alphanétisation

Dans cette partie nous cherchons à observer la relation entre le taux d'alphanétisation et l'indice culturel (obtenu à partir des notes des musées que nous avons scrapper de Google Maps<sup>3</sup> comme mentionné dans le Chapitre 1 section "Acquisition des données" ).

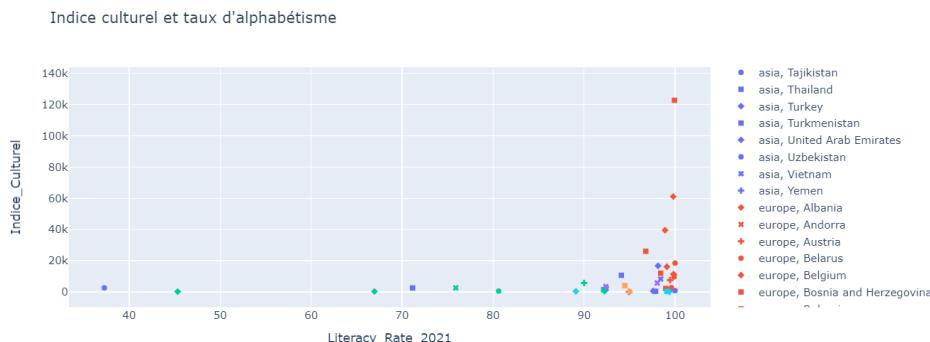


FIGURE 2.27 – Indice culturel et le taux d'alphanétisation par continent et pays

A partir de la figure 2.27 on peut voir que les pays européens ont majoritairement un indice culturel et un taux d'alphanétisation élevé. L'indice culturel est quant à lui relativement bas pour les pays de tous les autres continents. Enfin, il semble y avoir une disparité entre le taux d'alphanétisation des pays Africains, tout comme ceux de l'Asie.

Notons que ces données peuvent être biaisées pour certains pays car, pour certains, il manquait des musées que nous n'avons pas pu récupérer (ce problème est relatif à l'intermédiaire<sup>4</sup> que nous avons utiliser pour collecter les codes HTML à scrapper de Google Maps).

### — Impact du budget de l'éducation nationale sur le taux d'alphanétisation

A travers cette analyse 2.28 nous cherchons à évaluer le ratio entre le taux d'alphanétisation et le pourcentage du budget alloué à l'éducation nationale (% du PIB).

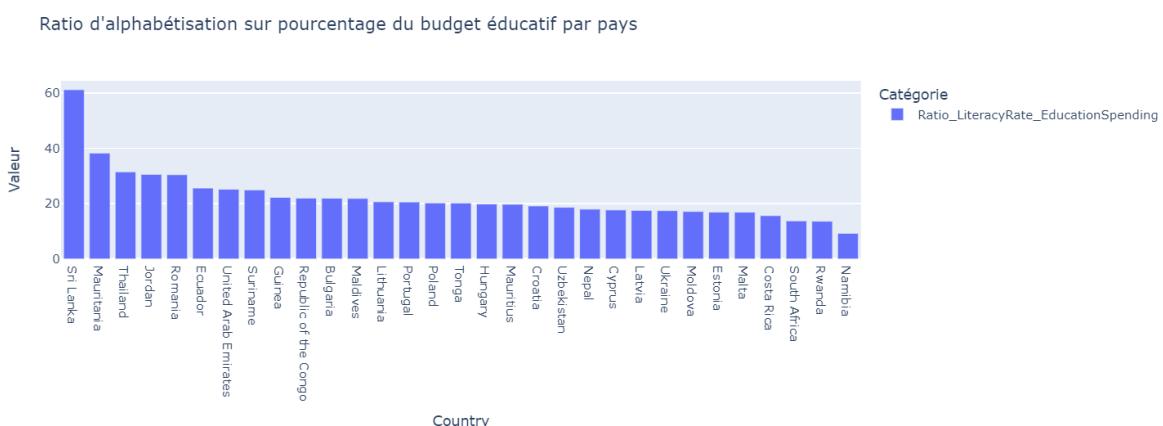


FIGURE 2.28 – Rapport entre le taux d'alphanétisation et le budget d'éducation nationale (%) par pays

3. [www.google.com/maps](http://www.google.com/maps)

4. [realtime.oxylabs.io](http://realtime.oxylabs.io)

Nous pouvons ainsi déduire que si le ratio est grand, tel que pour un pays comme le Sri Lanka, alors l’alphabétisation d’un pays repose principalement sur les institutions privées plutôt que publique (car le budget alloué par l’état est relativement petit).

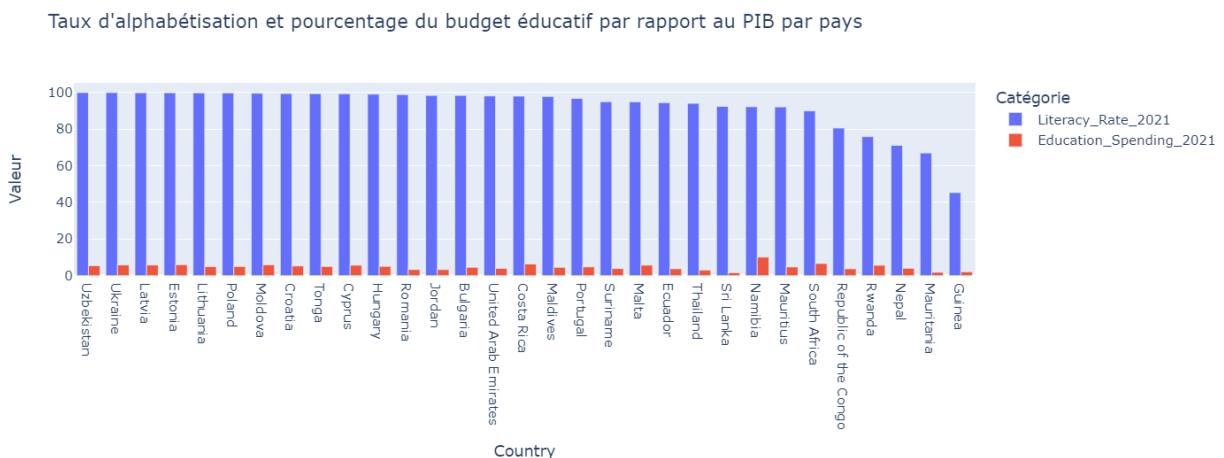


FIGURE 2.29 – Comparaison du taux d’alphabétisation et le budget d’éducation nationale (%) par pays

Dans le graphique ci-dessus (fig 2.29) nous affichons également ces proportions directement sans calculer le ratio.

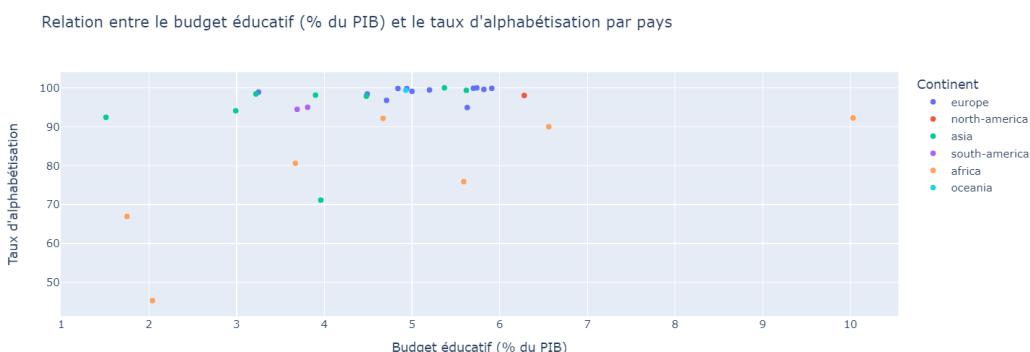


FIGURE 2.30 – Relation entre le taux d’alphabétisation et le budget d’éducation nationale (%) par continent

Cette dernière figure nous montre que le pourcentage du budget moyen alloué à l’éducation (% du PIB) par pays de chaque continent varie globalement entre 3% et 6%. De plus on observe que seul des pays d’Asie et d’Afrique descendent sous la barre des 90% du taux d’alphabétisation ; le budget éducatif alloué est également relativement faible, ce qui laisse penser que ces pays en question souffrent de problèmes prioritaires à gérer tel que répondre aux besoins élémentaires de leur population.

### — Analyse de données multivariée

Notre dernière analyse compare les continents par rapport à quatre paramètres : les prix nobels remportés, les universités figurant dans le top 500 des classements internationaux, le PIB et l’attraction des étudiants étrangers.

L'objectif étant de voir si certains continents se distinguent particulièrement par rapport à d'autres sur ces quatre paramètres ; paramètres que nous pensons avoir une importante influence sur le QI moyen des pays.

Notons que les données ont été normalisées de sorte à ce que les facteurs soient interprétables entre eux.

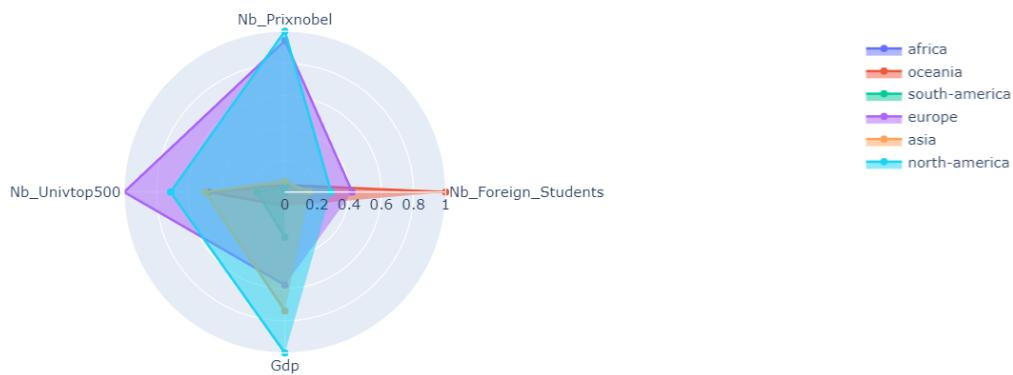


FIGURE 2.31 – Analyse de données multivariée

A partir de la figure 3.8, nous pouvons globalement dire que les pays d'Europe et d'Amérique du nord répondent le mieux aux facteurs analysés puisque qu'ils occupent une large surface sur le "Radar chart".

Cependant, il est intéressant de noter que les pays d'Océanie attirent le plus d'étudiants étrangers, alors que si on regarde leur nombre d'universités figurants dans le top 500 des meilleures universités, on les retrouverait derrière l'Europe et l'Amérique du nord. Cela peut indiquer que les motivations derrière ne sont pas directement liées à la qualité des études mais peut être à d'autres facteurs comme le coût des études.

## 2.3 Analyse en composantes principales (ACP)

Pour explorer les relations entre les caractéristiques de notre dataset tout en les visualisant dans un espace réduit de 2 à 3 dimensions, nous avons opté pour l'utilisation de l'Analyse en Composantes Principales (ACP).

Dans cette démarche, nous avons tout d'abord éliminé les observations pour lesquelles les valeurs des variables considérées étaient manquantes. Cette étape est cruciale pour garantir la qualité de notre analyse.

Ensuite, nous avons procédé à une standardisation des variables en les centrant sur leur moyenne et en les réduisant sur leur écart-type. Cette étape permet d'éliminer l'effet de l'échelle dans la construction des axes, une considération importante que nous avons identifiée lors de notre analyse préliminaire.

Une fois nos données prétraitées, nous avons utilisé la fonction ACP fournie par la bibliothèque Scikit-learn. Les résultats obtenus comprennent les axes factoriels avec leurs inerties expliquées ainsi que l'inertie totale expliquée.

<b>Dimension</b>		<b>Variance expliquée</b>	<b>% variance expliquée</b>	<b>% cum. var. expliquée</b>
<b>0</b>	1	4.638545	35.370842	35.370842
<b>1</b>	2	2.631939	20.069633	55.440474
<b>2</b>	3	1.444523	11.015091	66.455565
<b>3</b>	4	0.976599	7.446973	73.902538
<b>4</b>	5	0.941346	7.178157	81.080695
<b>5</b>	6	0.819529	6.249248	87.329943
<b>6</b>	7	0.476055	3.630119	90.960062
<b>7</b>	8	0.420887	3.209436	94.169498
<b>8</b>	9	0.257432	1.963028	96.132525
<b>9</b>	10	0.209781	1.599670	97.732195
<b>10</b>	11	0.158906	1.211724	98.943919
<b>11</b>	12	0.073860	0.563210	99.507129
<b>12</b>	13	0.064635	0.492871	100.000000

FIGURE 2.32 – Résultats ACP

Pour une visualisation plus approfondie, nous avons également représenté l'inertie expliquée par chacun des axes factoriels à l'aide d'un diagramme à barres.

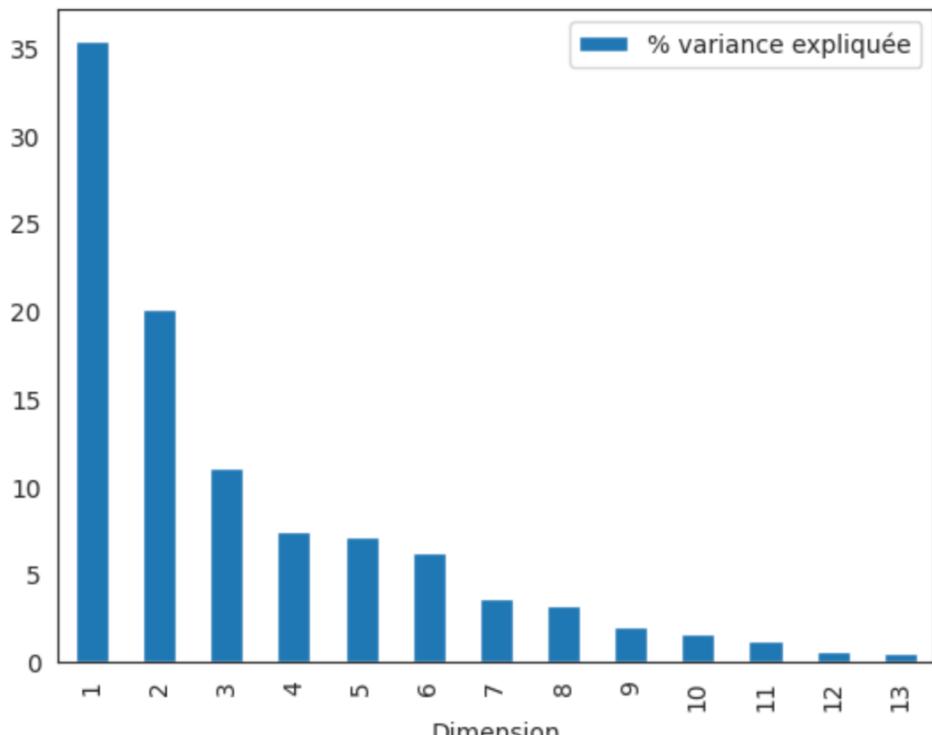


FIGURE 2.33 – Inerties expliquées par chacun des axes factoriels

### 2.3.1 Analyse des résultats

La première dimension explique la plus grande partie de la variance des données (**35.37%**), suivie par la deuxième dimension (**20.07%**). Ensemble, ces deux premières dimensions expliquent plus de la moitié de la variance totale (**55.44%**). Cela suggère que les deux premières dimensions sont importantes pour représenter les données de manière significative.

Les dimensions suivantes expliquent moins de variance individuellement, mais elles contribuent toujours à la compréhension globale des données. Par exemple, les cinq premières dimensions expliquent déjà plus de **80%** de la variance totale.

Les dimensions avec une faible variance expliquée peuvent être considérées comme moins importantes pour la représentation des données, mais elles peuvent toujours contenir des informations significatives, surtout si elles contribuent à la compréhension d'aspects spécifiques des données.

Pour une meilleure analyse des pays, nous les projetons sur le premier plan factoriel. Ainsi, la figure qui suit illustre cette projection. Dans ce graphique nous pouvons voir que les pays ayant des caractéristiques similaires sont proches les uns des autres. C'est le cas du deuxième axe factoriel. Il est à remarquer que les pays d'un même continent ne sont pas forcément proches les uns des autres. Ceci suggère qu'il y a des disparités au sein des pays d'un même continent.

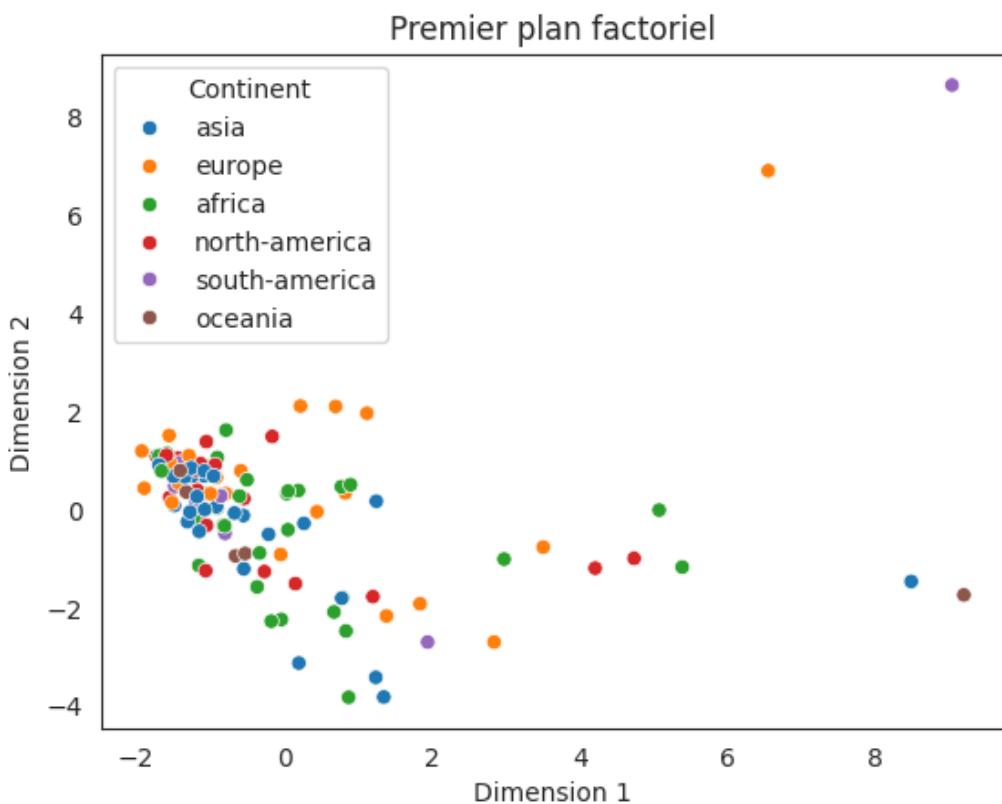


FIGURE 2.34 – Projection des pays sur le plan factoriel

Pour une meilleure analyse des variables et les relations, nous les projetons sur le premier plan factoriel. Ainsi, la figure qui suit illustre cette projection.

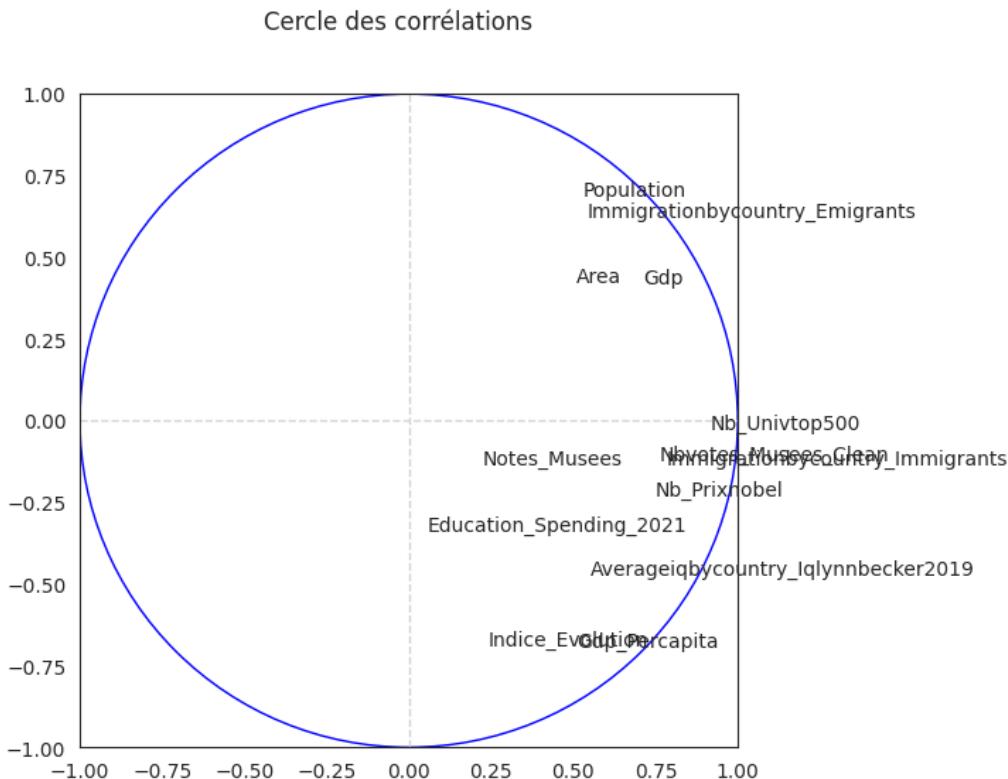


FIGURE 2.35 – Projection des variables sur le plan factoriel

Sur ce cercle de corrélation, nous pouvons voir dans le premier quadrant les variables corrélées que sont population et pourcentage émigrés, superficie et PIB. Ces dernières sont négativement corrélées par rapport aux variables du 4ème quadrant. Ceci signifie que les deux groupes de variables évoluent dans des directions inversées.

# Implémentations et expérimentations

---

Nous décrivons dans ce chapitre les différents cas d'application auxquels peuvent servir les données récoltées et analysées dans les chapitres précédents. Pour chaque cas d'application, nous présentons les modèles testés ainsi que les performances obtenues.

## 3.1 Modèles de Classification

Nous avons constaté que dans l'analyse exploratoire, les pays d'un même continent sont similaires par rapport aux caractéristiques que nous avons définis. Afin d'analyser cela de plus près, nous avons pensé à utiliser les caractéristiques des pays pour les classifier en continents. Ainsi, étant donné les caractéristiques d'un pays, le but est de prédire son continent. Pour ce faire, nous avons utilisé les modèles SVM, Régression logistique et XGBoost.

### 3.1.1 SVM

Le modèle SVM (Support Vector Machine) est un classificateur supervisé qui cherche à trouver l'hyperplan optimal séparant les classes dans un espace à haute dimension. L'idée principale est de maximiser la marge entre les points de données de différentes classes. Pour entraîner ce modèle, nous avons suivi les étapes suivantes :

1. Préparation des données : Normalisation des caractéristiques pour que chaque caractéristique ait une moyenne de 0 et une variance de 1.
2. Division des données : Séparation en ensembles d'entraînement et de test.
3. Entraînement du modèle : Utilisation de la fonction d'entraînement de SVM avec un noyau linéaire.
4. Prédiction : Application du modèle entraîné sur l'ensemble de test pour prédire les continents des pays.

Cependant, ce modèle a montré une performance limitée avec une précision de 30%, indiquant que les caractéristiques choisies ne sont pas suffisantes pour une classification précise par ce modèle.

### 3.1.2 Régression logistique

La régression logistique est un modèle de classification linéaire qui prédit la probabilité qu'un exemple appartienne à une classe donnée. La fonction sigmoïde est utilisée pour convertir la sortie linéaire en probabilité. Le pipeline classique pour ce modèle est :

1. Préparation des données : Normalisation des caractéristiques pour qu'elles soient comparables.
2. Division des données : Répartition des données en ensembles d'entraînement et de test.
3. Entraînement du modèle : Ajustement du modèle de régression logistique aux données d'entraînement.
4. Prédiction : Utilisation du modèle ajusté pour prédire les continents des pays dans l'ensemble de test.

La régression logistique a obtenu une précision de 34%, ce qui est légèrement meilleur que SVM mais reste faible pour des applications pratiques.

### 3.1.3 XGBoost

XGBoost (Extreme Gradient Boosting) est un modèle de boosting basé sur les arbres de décision. Il combine plusieurs arbres de décision faibles pour créer un modèle puissant et robuste. Les étapes pour entraîner ce modèle sont :

1. Préparation des données : Transformation et normalisation des caractéristiques.
2. Division des données : Séparation en ensembles d'entraînement et de test.
3. Entraînement du modèle : Utilisation de la fonction d'entraînement de XGBoost avec une recherche de paramètres optimaux.
4. Prédiction : Application du modèle sur l'ensemble de test pour prédire les continents.

XGBoost s'est révélé être le plus performant avec une précision de 56%. Bien que cette précision soit meilleure que les autres modèles, elle reste modeste.

### 3.1.4 Évaluation des Modèles

Nous avons comparé les performances des différents modèles en termes de précision sur l'ensemble de test. Les résultats sont résumés dans le tableau suivant :

Modèle	Précision (%)
SVM	30
Régression logistique	34
XGBoost	56

TABLE 3.1 – Précision des différents modèles de classification

### 3.1.5 Analyse des Résultats

Les résultats suggèrent que les caractéristiques utilisées ne sont pas suffisamment discriminantes pour permettre une classification précise des pays en fonction des continents. Même

le modèle XGBoost, bien que le plus performant, n'a atteint qu'une précision de 56%, ce qui est insuffisant pour des applications pratiques robustes. Cela indique que les caractéristiques choisies doivent être réévaluées et enrichies pour mieux capturer les différences significatives entre les continents.

## 3.2 Modèles de Clustering

Nous avons vu que la classification par continent n'est pas très adaptée dû à la variance entre les caractéristiques des pays au sein d'un même continent. Nous avons donc pensé à appliquer un clustering et voir quels pays sont considérés comme similaires. Pour cela nous avons choisis le modèle K-means et le clustering agglomératif.

### 3.2.1 K-means

L'algorithme K-means [1] est un algorithme non supervisé qui permet de partitionner un ensemble d'individus, ou observations, en k clusters en définissant des centres (centroïdes) pour chacun de ces clusters puis en appliquant un calcul de distance (le plus souvent euclidienne) entre les données du jeu de données en entrée.

Le pseudo code de l'algorithme est donné comme suit :

---

#### **Algorithm 1** Algorithme K-means

---

**Entrée :** Dataset (D), nombre de clusters (k)

**Sortie :** Clusters

- Choisir aléatoirement k points comme centroïdes (centre des clusters).
- **Répéter**
  - **Pour** chaque point dans D **faire** :
    - Calculer la distance euclidienne entre le point et les k centroïdes.
    - Affecter le point au cluster le plus proche.
    - Calculer les nouveaux centroïdes par rapport aux nouveaux points affectés.

**Jusqu'à** stabilisation des clusters.

**Renvoyer** Clusters

**Fin**

---

Une fois l'algorithme appliqué à l'ensemble de nos **données numériques**, nous appliquons la méthode du coude « K-needle » qui se base sur le principe d'inertie pour identifier le nombre de clusters à former.

L'inertie est la somme des distances au carré des points d'un cluster par rapport au centre

de ce dernier est donnée par la formule suivante :

$$\sum_{i=1}^p d^2(X_i, G_i^t)$$

où :

- $p$  : le nombre de clusters choisis par l'utilisateur,
- $X_i$  : l'ensemble de points du cluster  $i$ ,
- $G_i^t$  : le vecteur transposé du centre de gravité du cluster  $i$ .

Ainsi, en calculant l'inertie pour chaque valeur de  $k$  (nombre de clusters) entre 2 et 20, nous obtenons le graphique suivant 3.1 :

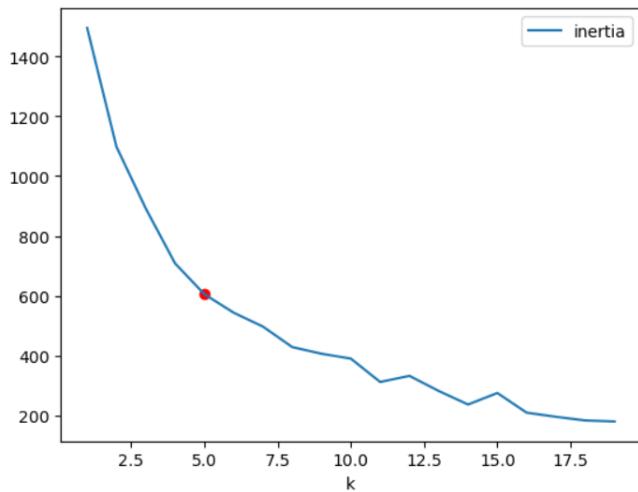


FIGURE 3.1 – Méthode du coude appliquée à nos données

Nous pouvons donc dire que les données devraient être réparties sur cinq clusters, notons que cette valeur de  $k$  se rapproche du nombre de continents.

Une fois la liste des pays appartenant à chaque cluster récupérée, nous les affichons sur une carte interactive comme le montre la figure 3.2.

## Regroupement des pays en clusters

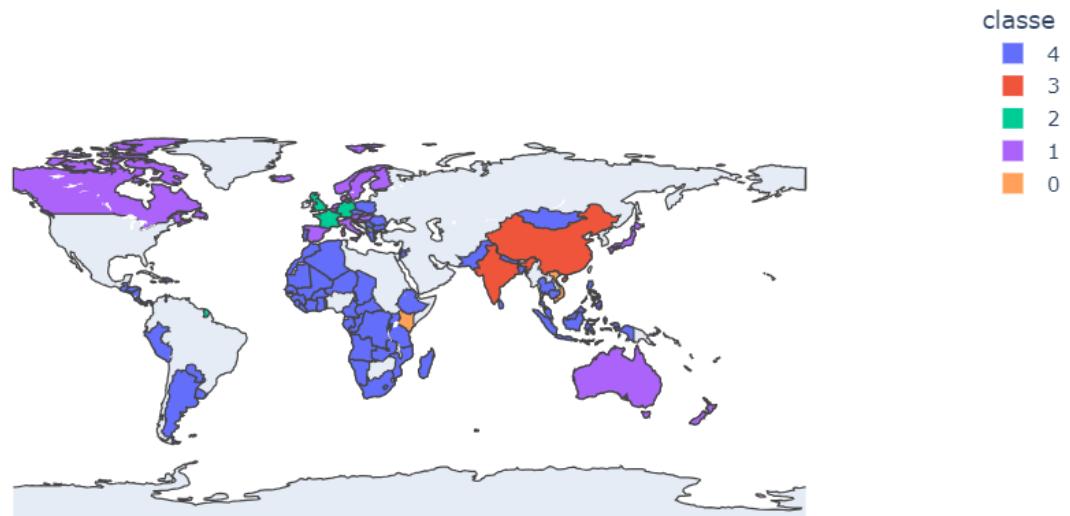


FIGURE 3.2 – Carte du monde avec des clusters de pays

On peut voir que la France, l'Allemagne et le Royaume-Uni ont été regroupés dans le même cluster ; ce qui paraît logique de par tous les facteurs qui les relient.

De plus, les pays Africains se sont tous vu attribuer la même classe, avec en plus quelque pays d'Asie et d'Amérique du Sud.

L'algorithme a aussi regroupé la Chine et l'Inde dans une même classe, et puis à considérer les autres pays tels que le Luxembourg, les Pays-Bas et le Danemark dans le cluster numéro 1 (violet) de la figure 3.2.

Un dernier cluster regroupe : le Kenya, le Laos, Malte et le Vietnam.

Nous terminons notre analyse en appliquant une ACP sur les données tel que chaque couleur correspond à un label attribué par K-means.

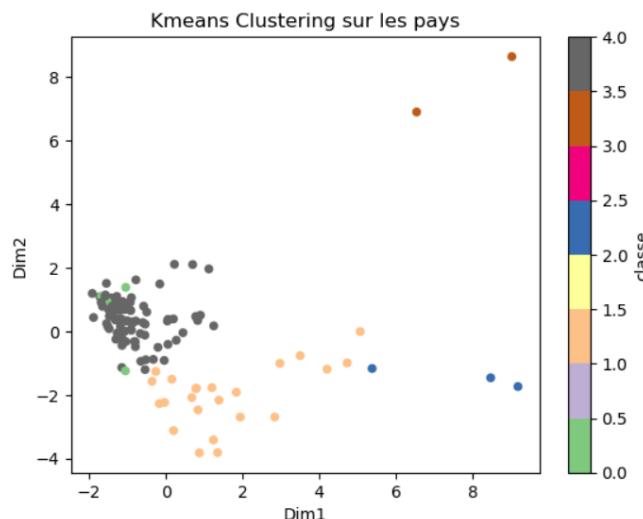


FIGURE 3.3 – Kmeans clustering sur les pays (ACP)

### 3.2.2 Clustering Agglomératif

Le clustering Agglomératif fait également partie de la famille d’algorithmes non supervisées pour la segmentation de données. Il repose sur une approche hiérarchique permet d’explorer la structure intrinsèque des données et de visualiser les relations de similarité à différentes échelles. Contrairement à des méthodes de clustering partitionnelles telles que K-means, qui nécessitent de spécifier le nombre  $k$  de clusters, le clustering agglomératif construit une hiérarchie de clusters en fusionnant progressivement les paires de clusters les plus similaires.

Notons que les mesures de similarité généralement utilisées sont la distance euclidienne, la corrélation de Pearson et la distance de Manhattan.

En appliquant cet algorithme sur nos **donnée numériques**, nous obtenons le dendrogramme suivant :

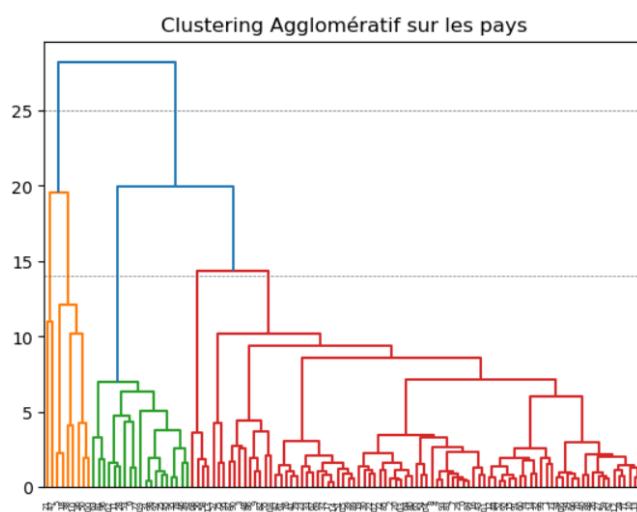


FIGURE 3.4 – Dendrogramme du clustering Agglomératif sur les pays

A partir de la figure 3.4, nous pouvons remarquer un saut important entre 2 et 4 classes, puis entre 5 et 6 classes. De ce fait, nous proposons de couper la hiérarchie à un de ces deux niveaux ; nous choisissons de la couper au second niveau de sorte à former cinq clusters et voir s’ils sont formés des mêmes pays que ceux des clusters de K-means.

Les résultats des clusters sont assez similaires à ceux produits par K-means, comme le montre la figure ci-dessous.

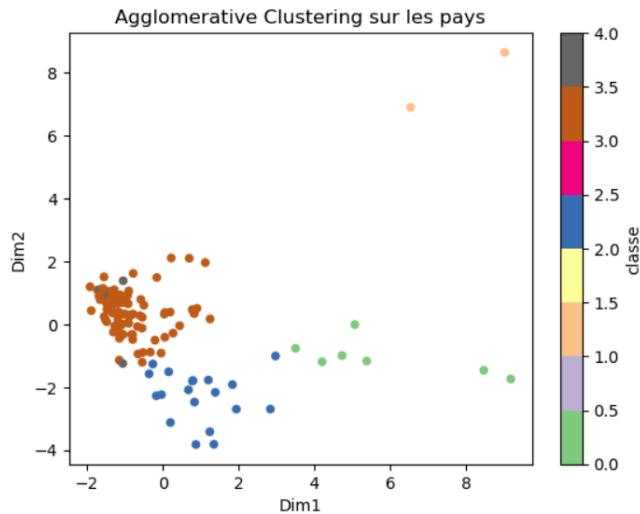


FIGURE 3.5 – Clustering Agglomératif sur les pays (ACP)

### 3.3 Modèles de Régression Linéaire

Une fois que nous avons analysé les différentes variables et leurs corrélations, nous en venons au modèle principal de notre étude qui consiste à prédire le QI moyen d'un pays étant donné ses caractéristiques. Il s'agit donc d'une régression. Pour ce faire, nous avons utilisé une régression linéaire et un réseau de neurones. Pour pouvoir les comparer, nous avons effectué une recherche en grille (grid search) pour chacun des modèles afin d'obtenir la meilleure combinaison des hyper-paramètres.

#### 3.3.1 Régression linéaire

Pour la régression linéaire, nous avons testé différentes combinaisons de régularisation et de normalisation des données. Plus précisément, nous avons testé les régularisations Ridge et Lasso, avec et sans normalisation des données (Standard Scaler). Les métriques utilisées pour évaluer les performances de nos modèles sont le coefficient de détermination ( $R^2$ ), l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne logarithmique (MSLE).

Le pipeline suivi pour la régression linéaire est le suivant :

1. **Préparation des données** : Division des données en ensembles d'entraînement et de test.
2. **Normalisation des données** : Application du Standard Scaler (ou non).
3. **Entraînement du modèle** : Entraînement des modèles de régression Ridge et Lasso.
4. **Évaluation des performances** : Calcul des métriques  $R^2$ , MAE et MSLE sur l'ensemble de test.

La meilleure combinaison de paramètres a été obtenue avec la régularisation Ridge et les données normalisées. Les résultats obtenus sont les suivants :

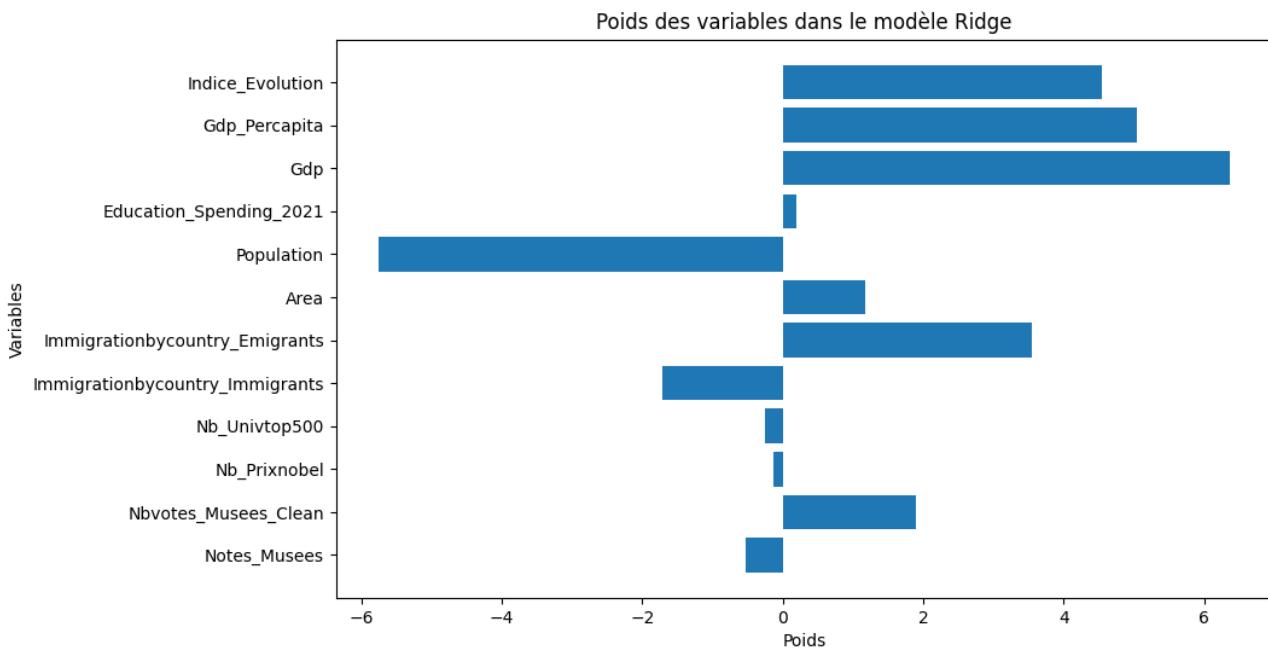


FIGURE 3.6 – Poids de certaines variables dans la prédiction

- Coefficient de détermination ( $R^2$ ) sur les données de test (Ridge) : 0.149
- Erreur absolue moyenne (MAE) sur les données de test (Ridge) : 9.741
- Erreur quadratique moyenne logarithmique (MSLE) sur les données de test (Ridge) : 0.017

### 3.3.2 Réseau de neurones

Pour le réseau de neurones, nous avons utilisé un réseau à 3 couches avec l'activation ReLU et l'optimiseur Adam avec un taux d'apprentissage (lr) de 0.001. Le batch size était de 32 et le nombre d'époques était de 10000, en prenant le meilleur modèle lors de l'entraînement. Les métriques d'évaluation étaient les mêmes que pour la régression linéaire :  $R^2$ , MAE et MSLE.

Le pipeline suivi pour le réseau de neurones est le suivant :

1. **Préparation des données** : Division des données en ensembles d'entraînement et de test.
2. **Normalisation des données** : Application du Standard Scaler.
3. **Configuration du modèle** : Définition de l'architecture du réseau de neurones avec 3 couches cachées et l'activation ReLU.
4. **Entraînement du modèle** : Entraînement du réseau de neurones avec l'optimiseur Adam (lr = 0.001) sur 10000 époques avec un batch size de 32.
5. **Évaluation des performances** : Calcul des métriques  $R^2$ , MAE et MSLE sur l'ensemble de test.

Les résultats obtenus avec le réseau de neurones sont les suivants :

- Coefficient de détermination ( $R^2$ ) sur les données de test : 0.055
- Erreur absolue moyenne (MAE) sur les données de test : 9.13
- Erreur quadratique moyenne logarithmique (MSLE) sur les données de test : 0.0128

Le graphique ci-dessous montre l'évolution de la perte (loss) en fonction des époques pour les ensembles d'entraînement et de test du réseau de neurones.

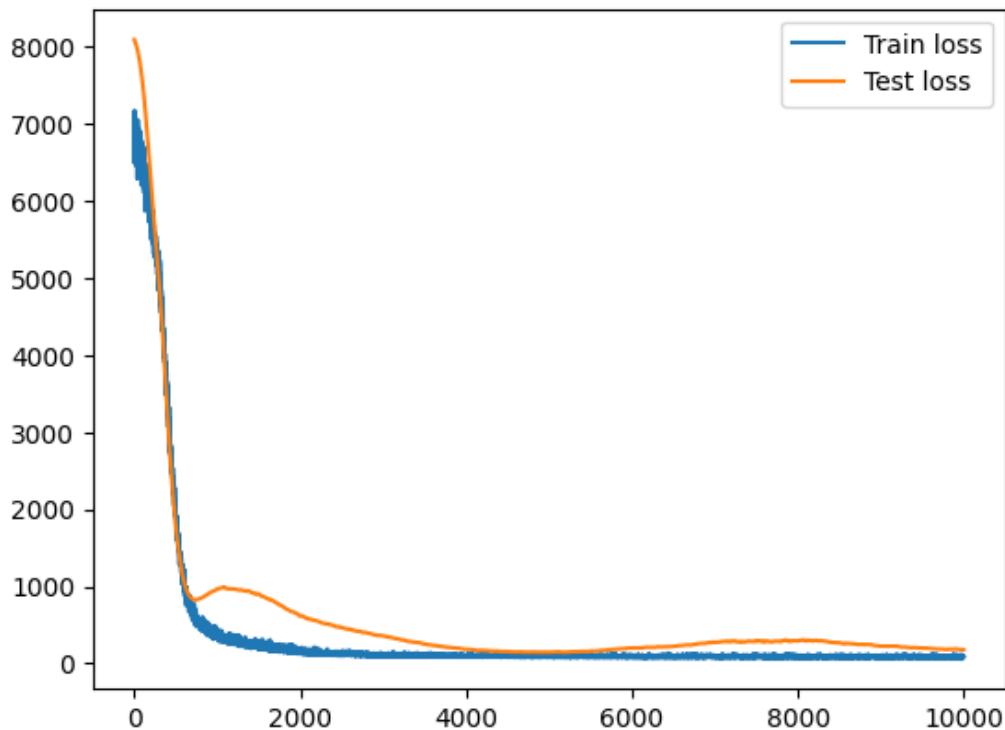


FIGURE 3.7 – Evolution de la loss

Pour expliquer certaines instances, nous avons utilisé Lime, (en anglais, Local Interpretable Model-agnostic Explanations) est un modèle local qui cherche à expliquer la prédiction d'un individu par analyse de son voisinage. A titre d'exemple, prenons la Belgique. La figure qui suit illustre le résultat retourné par Lime. Nous pouvons extraire les variables qui ont positivement contribué à ce score QI en l'occurrence le PIB, nombre d'universités dans le top 500 ou encore l'indice évolution ou encore les variables ayant négativement contribué à ce score que sont la surface et nombre de prix Nobel, cela suggère que la Belgique aurait pu avoir un meilleur score si elle avait plus de prix Nobel.

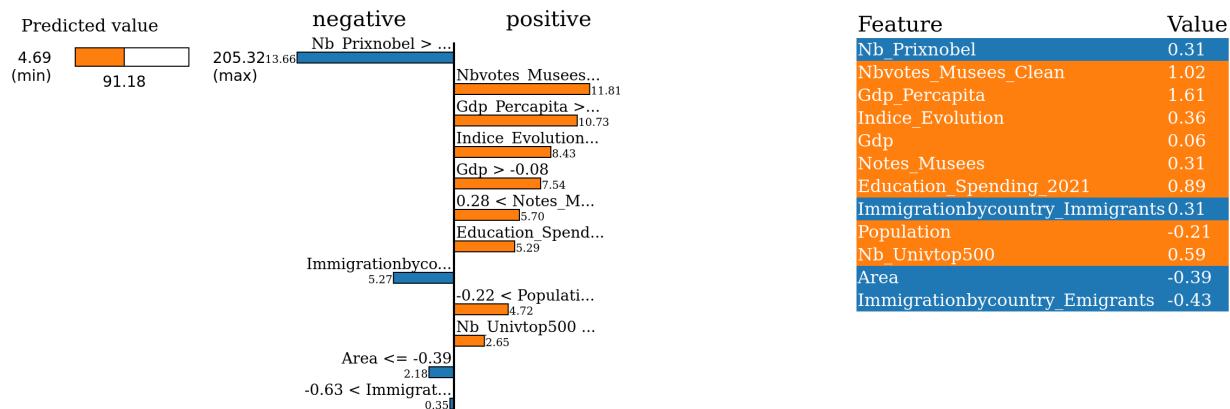


FIGURE 3.8 – Explications de la prédition du QI pour la Belgique

### 3.3.3 Comparaison des résultats

Modèle	$R^2$	MAE	MSLE
Régression Ridge	0.149	9.74	0.017
Réseau de neurones	0.055	9.13	0.0128

TABLE 3.2 – Comparaison des résultats entre les modèles de régression linéaire et de réseau de neurones.

Les résultats montrent que le réseau de neurones surpassé la régression linéaire en termes de coefficient de détermination ( $R^2$ ), d'erreur absolue moyenne (MAE) et d'erreur quadratique moyenne logarithmique (MSLE).

## 3.4 Modèles de Séries Temporelles

Nous terminons cette série d'analyses par une étude de séries temporelles. L'idée dans partie est de prédire l'indice évolution<sup>1</sup> des pays pour les N prochaines années. Pour ce faire nous calculons les indices évolution de chaque pays en exploitons le fait d'avoir des données de PIB allant de l'année 1960 jusqu'à 2022 pour la grande majorité des pays (la formule de calcul est donnée dans le chapitre 1 ainsi que dans la section 2.2.2). Les modèles utilisés sont ARIMA et Prophet que nous présentons dans ce qui suit.

### 3.4.1 ARIMA

l'AutoRegressive Integrated Moving Average (ARIMA) est l'un des modèles les plus utilisés pour modéliser et prévoir les séries temporelles. De par ses composantes, il se distingue par sa

1. mesure que nous avons calculé dans les précédents chapitres, qui estime à quel point un pays évolue rapidement.

capacité à capturer à la fois les tendances, les saisons et les comportements auto-régressifs des données. Les composantes en question sont les suivantes :

- **AR (AutoRégression)** : Cette composante capture les relations linéaires entre les observations successives dans la série temporelle. Elle est basée sur l'hypothèse selon laquelle les valeurs futures de la série dépendent linéairement de ses valeurs passées.
- **I (Intégration)** : Cette composante est utilisée pour rendre la série temporelle stationnaire en différenciant les observations. Notons que la stationnarité est une propriété importante car elle garantit que les propriétés statistiques de la série restent constantes dans le temps.
- **MA (Moyenne Mobile)** : Cette composante modélise les erreurs de prédiction en termes d'une combinaison linéaire des erreurs de prédiction passées. Elle permet ainsi de réduire le bruit présent dans les séries temporelles.

## Analyse et traitement des données

Dans le cadre de notre étude, nous avons procédés à une analyse de données puis à leur traitement afin qu'ils soient conformes aux critères du modèle ARIMA à savoir :

- La stationnarité des données, nos données doivent être transformés en une série stationnaire ; c'est à dire faire en sorte que la moyenne et la variance sont constantes dans le temps. Nous utilisons le test Dickey-Fuller (ADF) afin de vérifier cette propriété.
- L'absence de tendance dans la série, dans le cas où une tendance est présente, nous faisons en sorte de la différencier afin de la rendre linéaire.
- l'absence de saisonnalité ; à savoir la variations cycliques à des périodes régulières.
- Les données doivent avoir une variance constante dans le temps.

Notons que nous avons traité les valeurs manquantes en les remplaçant par la valeur précédente, nous ne voulions pas prendre le risque de les remplacer par la moyenne, car auquel cas nous retrouverions un pique anormal dans la série.

Pour ce qui est des analyses, nous sélectionnons les données de la France et effectuons les expérimentations suivantes.

Dans un premier temps, nous affichons la série temporelle avec en plus un *rolling* (une fenêtre glissante de taille égale à cinq) sur la moyenne et l'écart type.

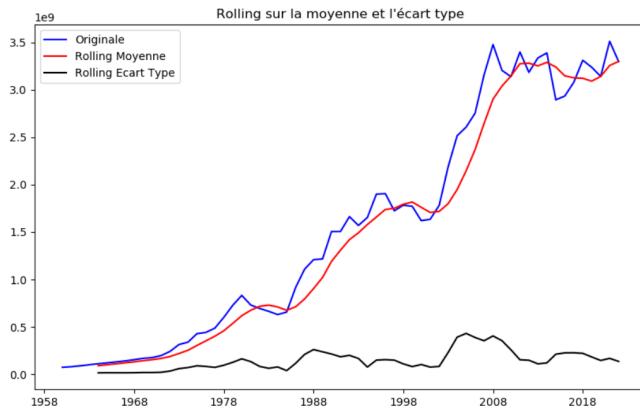


FIGURE 3.9 – Série temporelle de l’indice évolution de la France avec un rolling sur cinq ans

Nous remarquons que l’indice évolution est très grand, ce qui peut rendre la tâche de *forecasting* extrêmement difficile au modèle ; car la plage de valeurs possible est trop grande. Pour y remédier, nous appliquons le logarithme décimal ( $\log_{10}$ ) aux données.

De plus, nous remarquons une tendance, or nous avons préciser dans les critères précédents d’ARIMA qu’il fallait l’éliminer. Pour ce faire, il suffit de soustraire des données la moyenne obtenue lors du fenêtrage. Nous obtenons la série temporelle suivante :

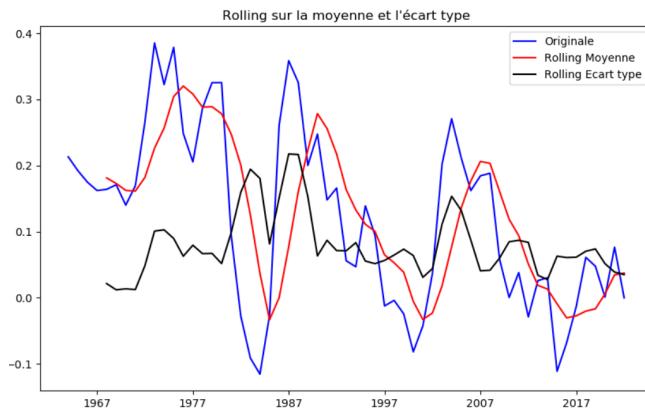


FIGURE 3.10 – Série temporelle de l’indice évolution de la France sans tendance

De plus, nous affichons dans la figure 3.11 une analyse des composantes temporelles de notre série de données log-transformées.

Nous observons des données de seasonality et residuals à 0, cela est très probablement dû au manque de données, il est difficile pour l’algorithme de décomposition de détecter des tendances et des motifs saisonniers significatifs.

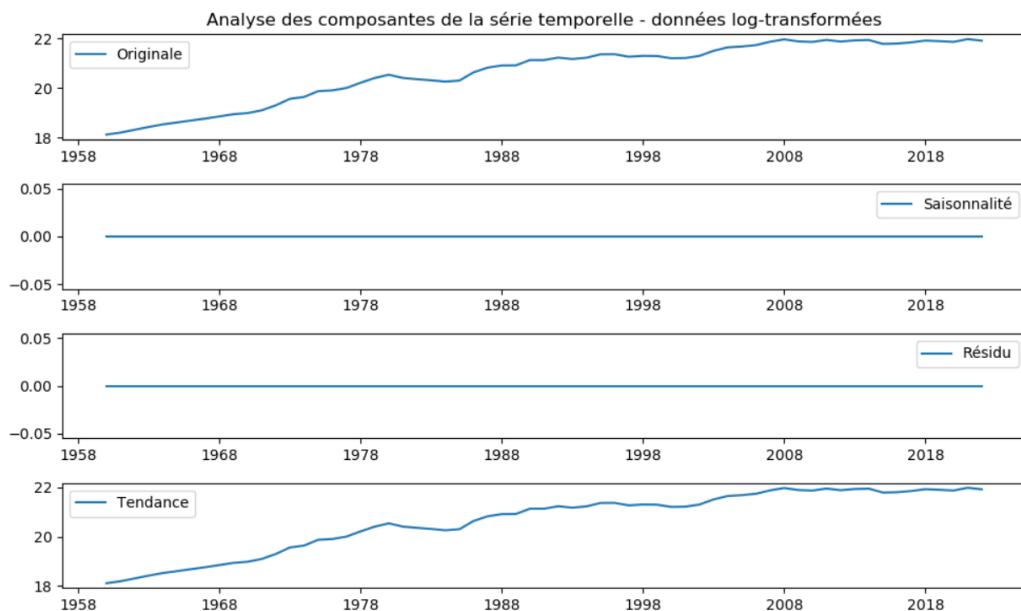


FIGURE 3.11 – Analyse des composantes temporelles - ARIMA

Pour vérifier la propriété d'absence de tendance dans les données, nous appliquons le test de Dickey Fuller, et obtenons les valeurs décrites dans le tableau 3.3.

Données	Valeurs
Test Statistic	-3.294298
p-value	0.015135
#Lags Used	1.0
Nombre d'instances	57.0
Critical Value (1%)	-3.550670
Critical Value (5%)	-2.913766
Critical Value (10%)	-2.594624

TABLE 3.3 – Résultats du test de Dickey Fuller

- **Définitions et mise en contexte :** Le test de Dickey-Fuller est principalement utilisé pour tester la stationnarité d'une série temporelle. Cependant, nous l'utilisons dans ce contexte pour vérifier l'absence de tendance étant donné que la stationnarité implique souvent ce dernier critère, car une série avec une tendance déterministe (linéaire ou autre) n'est généralement pas stationnaire.
  - Test Statistic : C'est la statistique du test calculée pour le test de Dickey-Fuller. Cette valeur est comparée aux valeurs critiques pour déterminer si la série est stationnaire.
  - p-value : C'est la valeur p associée à la statistique du test. Une p-value faible (généralement inférieure à 0.05) indique que l'hypothèse nulle de non-stationnarité peut être rejetée, suggérant que la série est stationnaire. Nous définissons les paramètres évalués par le test par ce qui suit.

- #Lags Used : est le nombre de retards utilisés dans le test de Dickey-Fuller ; ce sont des termes supplémentaires utilisés pour éliminer l'autocorrélation dans les résidus.
- Nombre d'instances : est le nombre d'observations dans la série temporelle.
- Critical Values : sont les valeurs critiques pour différents niveaux de confiance (1%, 5%, et 10%).

L'idée est de comparer la statistique du test aux critical values pour déterminer si l'hypothèse nulle peut être rejetée.

A partir du tableau 3.3, nous pouvons observer que la statistique de test -3.294298 est plus faible que la valeur critique à 5%, qui est à -2.913766, mais pas que celle à 1%, qui est à -3.550670. Cela signifie que nous pouvons rejeter l'hypothèse nulle de non-stationnarité au niveau de signification de 5%, mais pas au niveau de 1%.

Pour ce qui est de la p-value, sa valeur est à 0.015135 ; soit est inférieure à 0.05, ce qui indique également que nous pouvons rejeter l'hypothèse nulle au niveau de signification de 5%.

Ainsi, nous pouvons conclure que la série est probablement stationnaire au niveau de signification de 5%, mais il y a une certaine incertitude si nous utilisons un seuil de 1%.

Nous terminons l'étude des données temporelles par l'analyse de l'Autocorrélation (ACF) et la Partial Autocorrélation (PACF).

- **Définitions et mise en contexte :** La ACF mesure la corrélation entre une série temporelle et ses propres valeurs décalées dans le temps, elle indique ainsi dans quelle mesure les valeurs de la série temporelle à différents moments sont liées les unes aux autres. Pour ce qui est de la PACF, elle permet de distinguer plus clairement les relations directes entre les valeurs de la série temporelle à différents, en plus de ce que fait la ACF, la PACF contrôle l'effet des observations situées entre ces deux intervalles (décalages).

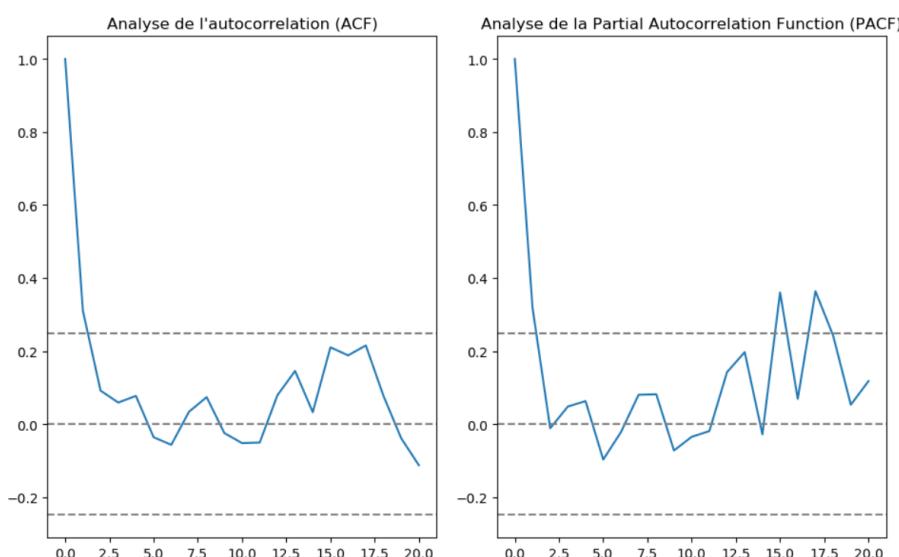


FIGURE 3.12 – Analyse de l'Autocorrélation (ACF) et la Partial Autocorrélation (PACF)

A partir des graphiques de la figure 3.12, nous déduisons que la valeur de l'ACF devrait être égale à 5 et celle de la PACF à 2 ; cela correspond à la première valeur des abscisses pour laquelle l'axe des coordonnées s'annule.

## Application des modèles

Pour ce qui est des modèles, nous testons dans un premier temps le modèle AR et MA séparément, avec les valeurs de ACF et PACF précédemment données. Nous obtenons les graphiques de la figure suivante.

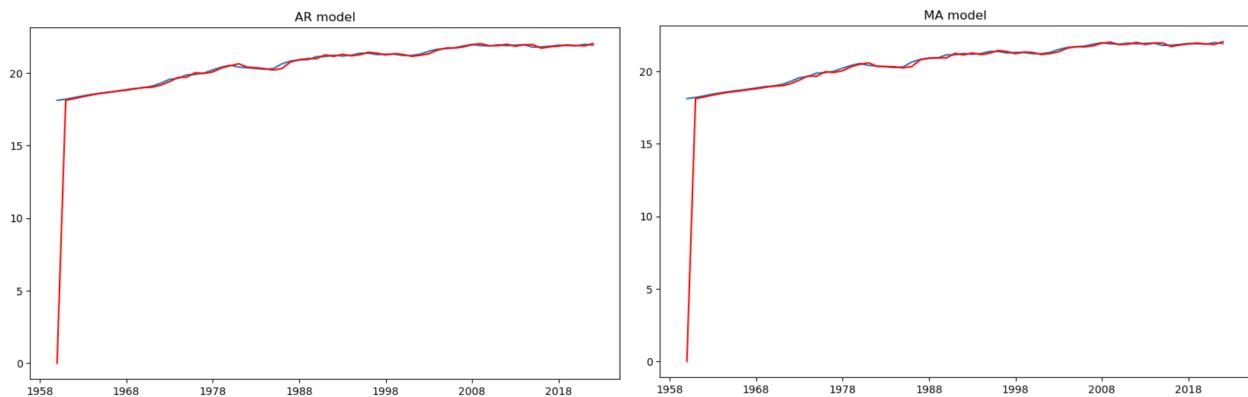


FIGURE 3.13 – Test du modèle AR et du modèle MA

Nous testons ensuite le modèle ARIMA (toujours avec les valeurs de ACF et PACF précédemment données) et obtenons le graphique de la figure 3.14.

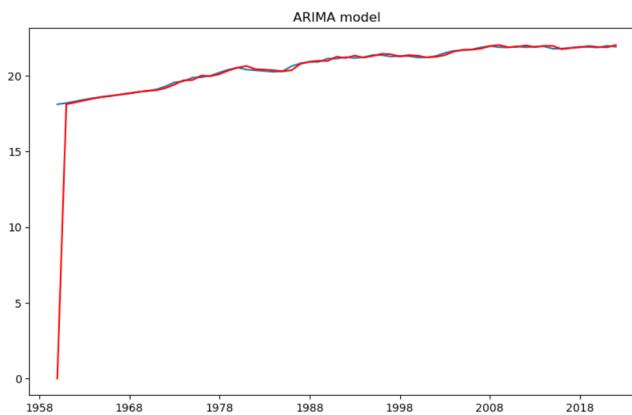


FIGURE 3.14 – Test du modèle ARIMA

Pour l'évaluation des modèles, nous utilisons le score de Residual Sum of Squares (RSS). C'est une mesure de l'erreur de prédiction du modèle ARIMA. Plus l'erreur est faible, plus ça indique que le modèle s'ajuste correctement aux données. Cette métrique est donnée par la formule suivante :

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Avec :

- $y_i$  : représente les valeurs réelles de la série temporelle.
- $\hat{y}_i$  : représente les valeurs prédites par le modèle.
- $n$  : représente le nombre total d'observations dans la série temporelle.

Les scores RSS obtenus sont égales respectivement à 328.80 et 328.85 et 328.77 pour AR, MA et ARIMA ; ce qui montre que ARIMA est meilleur que les deux autres, nous le choisissons pour le reste de nos expérimentations dans cette partie.

Nous terminons par afficher les prédictions du modèle (ARIMA) sur la série temporelle de la France avec un intervalle de confiance de 95%, comme le montre la figure ci-dessous. Nous remarquons que la largeur de l'intervalle de confiance augmente fortement lorsque le modèle doit prédire une valeur à une année sur laquelle il n'a pas été ajusté.

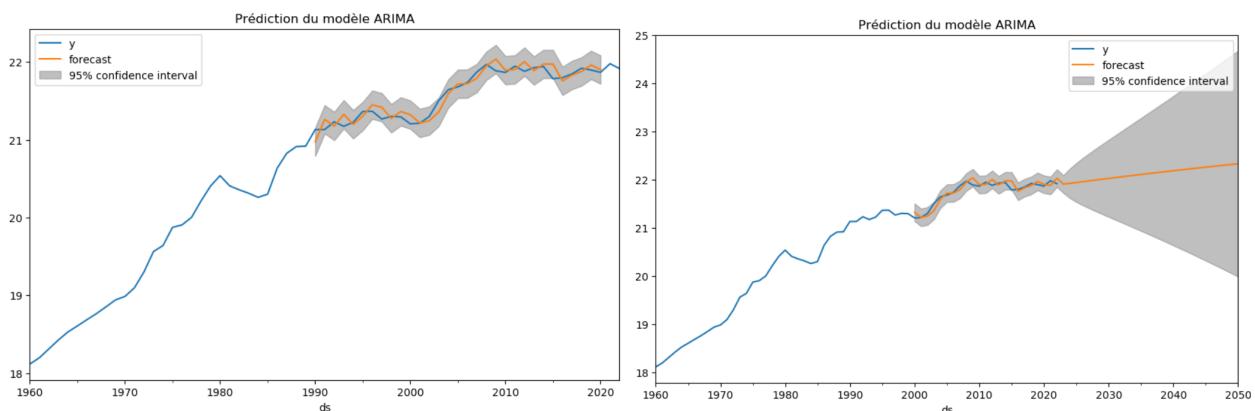


FIGURE 3.15 – Prédiction du modèle ARIMA

### 3.4.2 Prophet

Le modèle Prophet<sup>2</sup> utilisé est proposé par l'entreprise Meta<sup>3</sup> et est décrit dans le papier de recherche [2] intitulé « Forecasting at scale ». A travers cet article, les auteurs visent à résoudre les défis associés à la production de prévisions fiables et de haute qualité, surtout lorsque les séries temporelles sont nombreuses et que les individus qui les manipulent ne sont pas experts en modélisation de séries temporelles.

Nous affichons cette fois ci les données relatives à la France mais avec le *forecasting* de Prophet. De plus, nous présentons la tendance mensuelle et annuelle des données. Nous obtenons ainsi le graphique ci-dessous.

2. <https://facebook.github.io/prophet/>  
 3. <https://opensource.fb.com/>

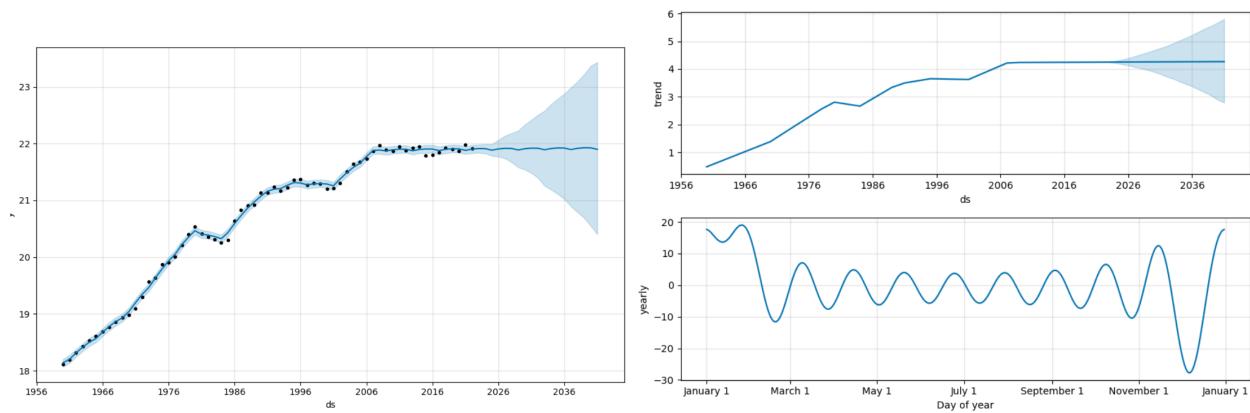


FIGURE 3.16 – Test du modèle Prophet

Nous justifions la tendance mensuelle nulle par le fait que le modèle ne possède que les données du 01 Janvier de chaque année.

Enfin, nous comparons les prédictions des indices d'évolution entre la France et le Singapour et observons, à partir de la figure 3.17, que cet indice semble se stabilisé pour la France, contrairement au Singapour qui semble croître d'année en année.

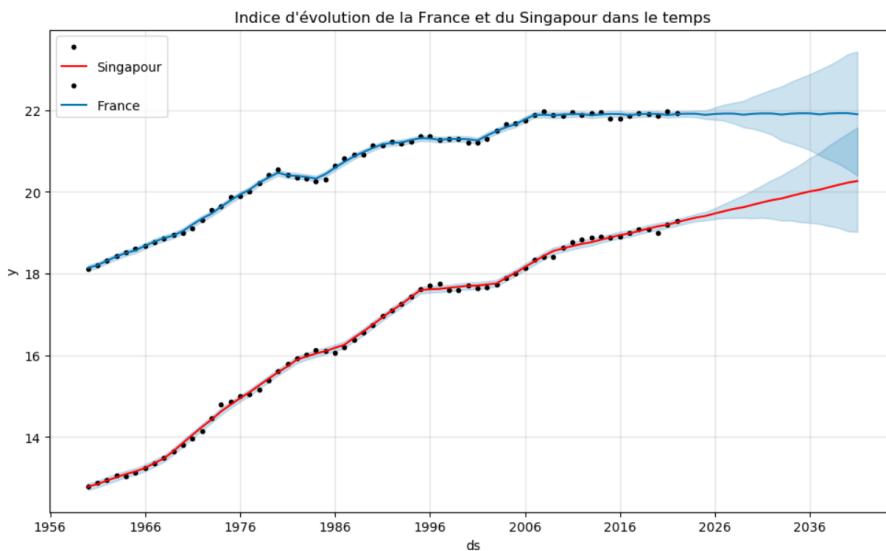


FIGURE 3.17 – Test de Prophet - comparaison entre la France et le Singapour

Notons que les données sur lesquels est appliqué Prophet sont également mises au Log10 ; pour garantir une meilleure stabilité du modèle.

# Mise en production de la solution

A travers ce chapitre, nous présentons les fonctionnalités de notre tableau de bord ; soit l'interface utilisateur qui rend l'accès à notre solution plus convivial. De plus, nous expliquons les étapes de déploiement de notre application.

## 4.1 Présentation du tableau de bord

Nous affichons dans ce qui suit les différentes sections qui constituent notre application. Notons que la grande majorité des graphiques sont interactifs et sont tous reliés aux données et modèles utilisés en **temps réel**.

### 4.1.1 Home : Analyse exploratoire des données

Dans cette partie nous affichons les axes d'analyses détaillés dans le chapitre 2. Notons que d'autres graphiques sont disponibles sur les fichiers notebook, sauf que nous avons estimé qu'il n'était pas assez pertinent de les afficher.

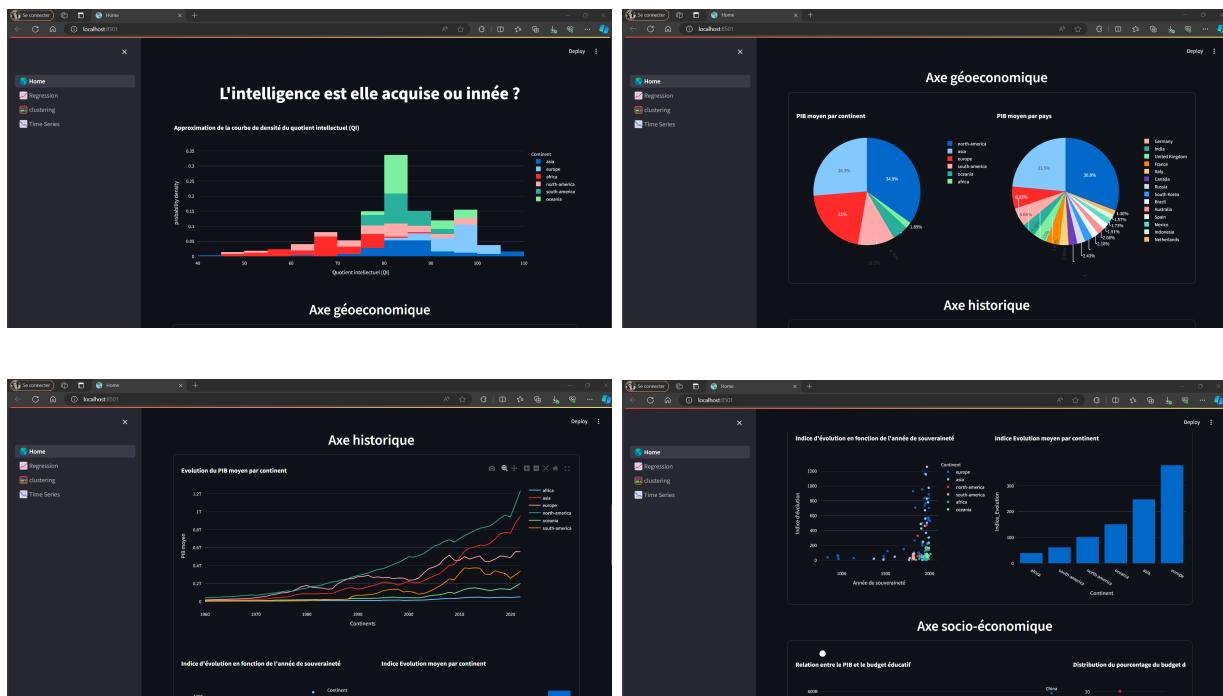


FIGURE 4.1 – EDA - Présentation des axes d'analyses (1)

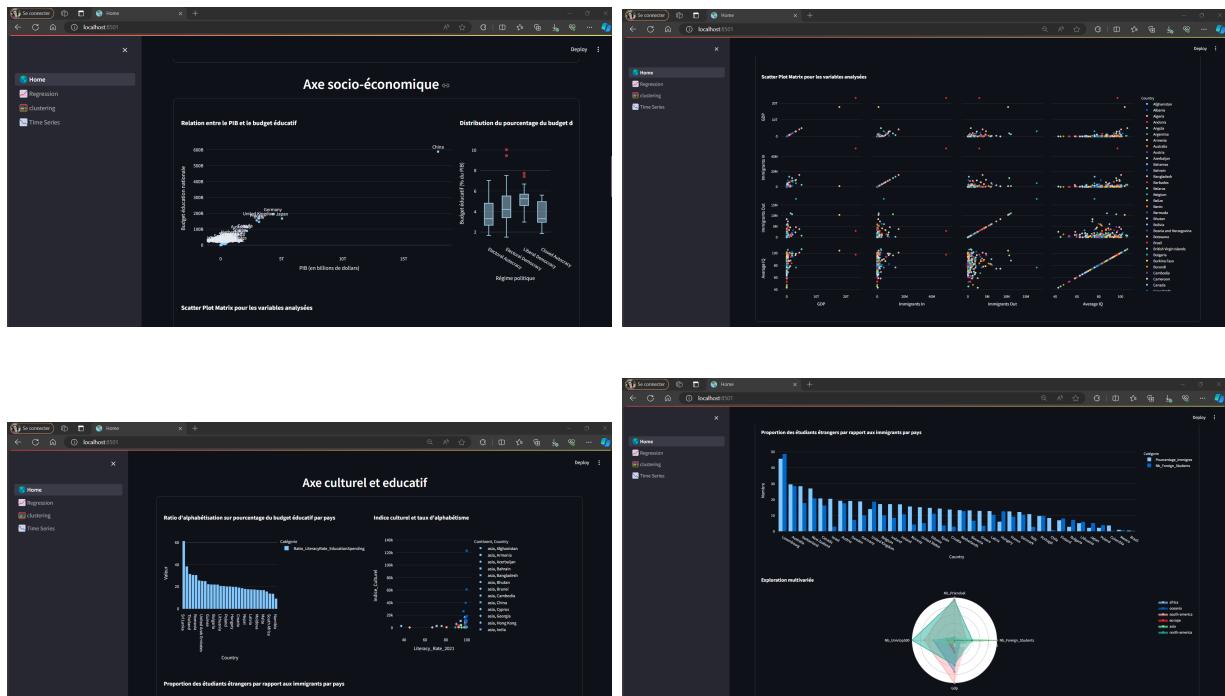


FIGURE 4.2 – EDA - Présentation des axes d'analyses (2)

### 4.1.2 Regression : Prédiction du QI

Afin de tester notre modèle pour prédire le QI moyen d'un pays, il suffit à l'utilisateur de préciser le nom du pays qu'il souhaite analyser, et l'application lui affiche le QI prédict *versus* le QI réel avec en plus le nombre de points de QI en plus/ou en moins par rapport à la vraie valeur. Aussi, nous avons intégrer une partie d'explicabilité grâce au module LIME qui affiche les attributs ayant participé le plus (positivement/négativement) à la prédiction du score pour le pays sélectionné.

Notons que nous avons utilisé en back-end notre modèle déjà entraîné sur une partie du jeu de données ; afin d'éviter à l'utilisateur un long temps d'attente.

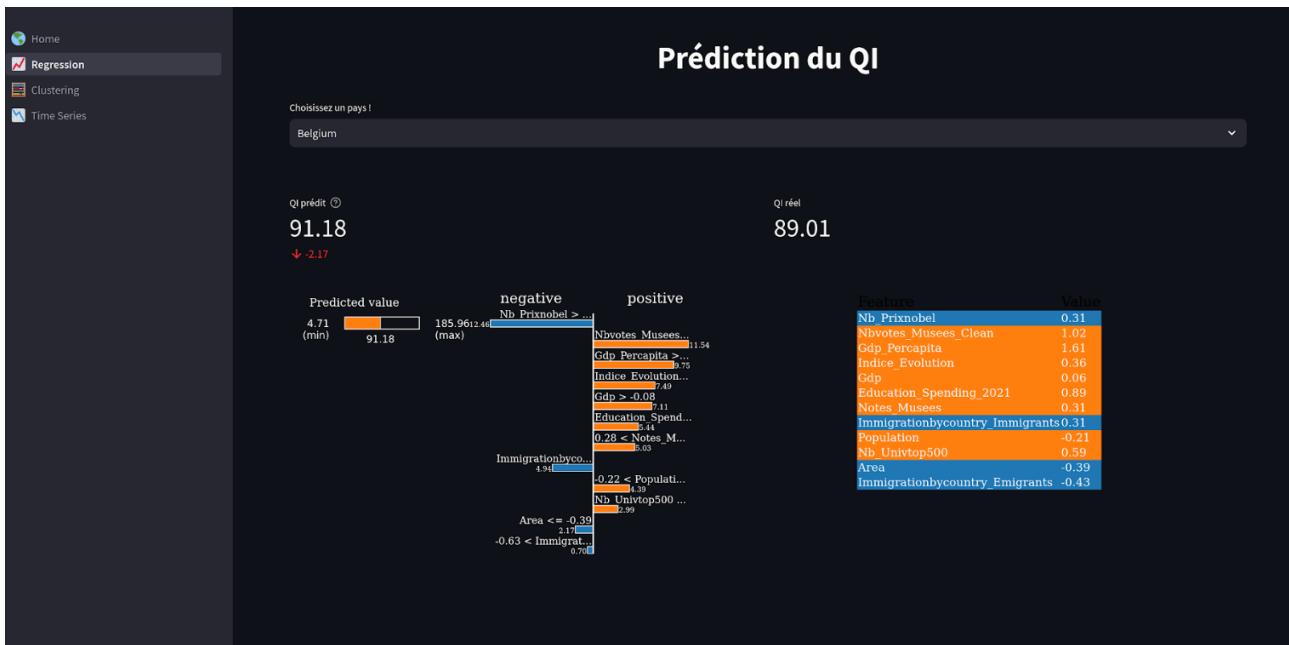


FIGURE 4.3 – Régression - Prédiction du QI

#### 4.1.3 Clustering : Regroupement des pays en clusters

Nous affichons dans cette partie la carte du monde avec une couleur attribuée à chaque pays en fonction de la classe à laquelle l'algorithme K-means l'a affecté (voir le chapitre 3 pour plus de détails sur l'analyse de cette carte).



FIGURE 4.4 – Clustering - Regroupement des pays similaires en classes

#### 4.1.4 Time Series : Prédiction de l'indice évolution dans le temps

L'application propose de comparer deux pays entre eux pour voir l'évolution possible de l'indice évolution de chacun ; pour rappel, l'indice évolution est une mesure que nous avons définis, qui estime à quel point un pays évolue rapidement.

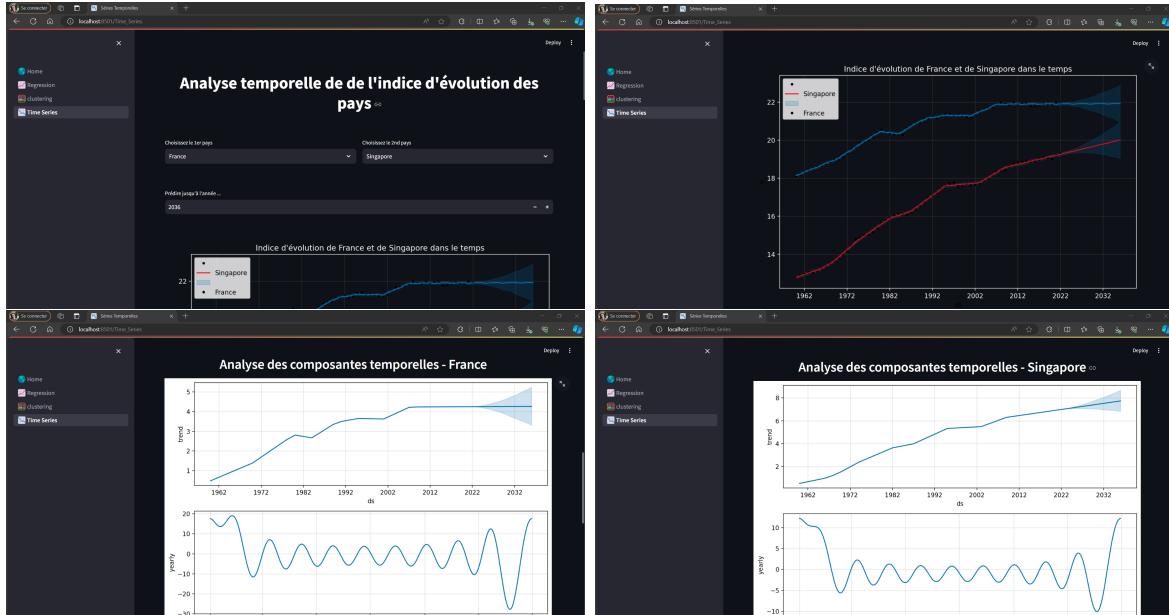


FIGURE 4.5 – Analyse des séries temporelles - Prédiction de l'indice évolution dans le temps

Pour ce faire, il suffit à l'utilisateur de choisir les deux pays qu'il souhaite comparer et préciser jusqu'à quelle année appliquer la prédiction et l'application lui retourne les graphiques associés. De plus nous proposons une analyse des composantes temporelles des deux séries, à savoir la tendance annuelle et mensuelle.

Notons que dans pour tous les pays, nous pouvons remarquer une tendance mensuelle nulle car le modèle ne possède que les données du 01 Janvier de chaque année.

## 4.2 Déploiement

Le déploiement de notre tableau de bord codé en Streamlit a été réalisé de manière méthodique et structurée, en utilisant divers outils et services pour garantir son accessibilité et sa fiabilité. Notre code est entièrement disponible en accès libre sur GitHub, hébergé dans le référentiel <https://github.com/ahmedmokeddem/dalas>.

#### 4.2.1 Utilisation de Docker pour la Conteneurisation

Nous avons choisi d'utiliser Docker pour la conteneurisation de notre application. Voici le contenu de notre fichier `Dockerfile` :

```

FROM python:3.10-slim

WORKDIR /app

RUN pip3 install torch torchvision torchaudio \
--index-url https://download.pytorch.org/whl/cu118
RUN python -m pip install "prophet==1.1.2"
RUN pip3 install plotly
RUN pip install numpy
RUN pip install streamlit
RUN pip3 install scikit-learn
RUN pip install "holidays==0.24"
RUN pip install lime
COPY ./dashboard/

EXPOSE 8501

HEALTHCHECK CMD curl --fail http://localhost:8501/_stcore/health

CMD ["streamlit", "run", "dashboard/_Home.py", \
"--server.port=8501", "--server.address=0.0.0.0"]

```

Ce Dockerfile définit un environnement Python, installe les dépendances requises, copie le contenu de notre répertoire local dans le conteneur, expose le port 8501 pour Streamlit, définit un point de contrôle de santé et spécifie la commande pour exécuter notre tableau de bord.

#### 4.2.2 Construction de l’Image Docker et Exécution du Conteneur

Pour construire l’image Docker, nous avons exécuté la commande suivante dans le répertoire de notre projet :

```
docker build -t nom_image .
```

Une fois l’image construite, nous avons exécuté le conteneur à l’aide de la commande suivante :

```
docker run -p 8501:8501 nom_image
```

Nous avons déployé ce conteneur sur une version communautaire de Streamlit, dont le site du tableau de bord déployé est `mokeddem-said.streamlit.app/`.

# Conclusion

---

La réalisation de ce projet nous a permis de développer une application sous forme de tableau de bord, qui représente différents axes d'analyses et plusieurs cas d'applications aux caractéristiques de pays. Nous avons pour cela suivi une méthodologie qui regroupe les principales étapes d'un projet de Data Science, allant de la collecte des données, à leurs traitement et analyse jusqu'à l'application de modèles d'apprentissage supervisés et non supervisés.

A travers cette étude, nous avons pu conclure que l'intelligence n'est finalement ni innée ni acquise mais dépend de ces deux facteurs à la fois.

Comme perspectives futures, il serait pertinent d'intégrer davantage de facteurs à explorer tel que l'analyse des tweets concernant le système éducatif de chaque pays. Enfin, il serait intéressant de déployer notre solution Docker sur un serveur web indépendant ; autre que streamlit.

# Bibliographie

---

- [1] Jiawei HAN, Micheline KAMBER et Jian PEI, *Data Mining: Concepts and Techniques*, 3rd, Waltham, MA, USA : Morgan Kaufmann, 2012.
- [2] Sean J TAYLOR et Benjamin LETHAM, « Forecasting at scale », in : *PeerJ Preprints* 5 (2017), DOI : 10.7287/peerj.preprints.3190v2, URL : <https://doi.org/10.7287/peerj.preprints.3190v2>.