Analyse de données

Data Mining

Angéline Marc, Thomas Huguenel
16/11/2024

Table des matières

Table des illustrations

Table des tableaux

Introduction

Le but de ce projet était de sélectionner un dataset et de faire une analyse des données grâce à plusieurs méthodes que nous avons étudiées en cours. Pour cela, nous avons choisi le jeu de données Gapminder disponible sur Kaggle.

Traitement des données

Notre dataset est composé de plusieurs colonnes :

- Country : Contient le nom de tous les pays présents
- Continent : Fait référence au continent sur lequel se situe le pays.
- Year : Année à laquelle correspond les données récoltées (de 1998 à 2018)
- Life_exp : Il s'agit de l'espérance de vie moyenne de la population du pays
- Hdi_index : Représente l'indice de développement humain (entre 0 et 1)
- Co2_consump : Désigne la quantité de CO2 émise pour une année en tonnes et par personne
- Gdp : Concerne le produit intérieur brut par habitant en dollars
- Services : Correspond au pourcentage de personnes travaillant dans le secteur tertiaire

Nous avons dû commencer notre analyse par du nettoyage de données, car certaines valeurs ne sont pas renseignées dans le jeu de données.

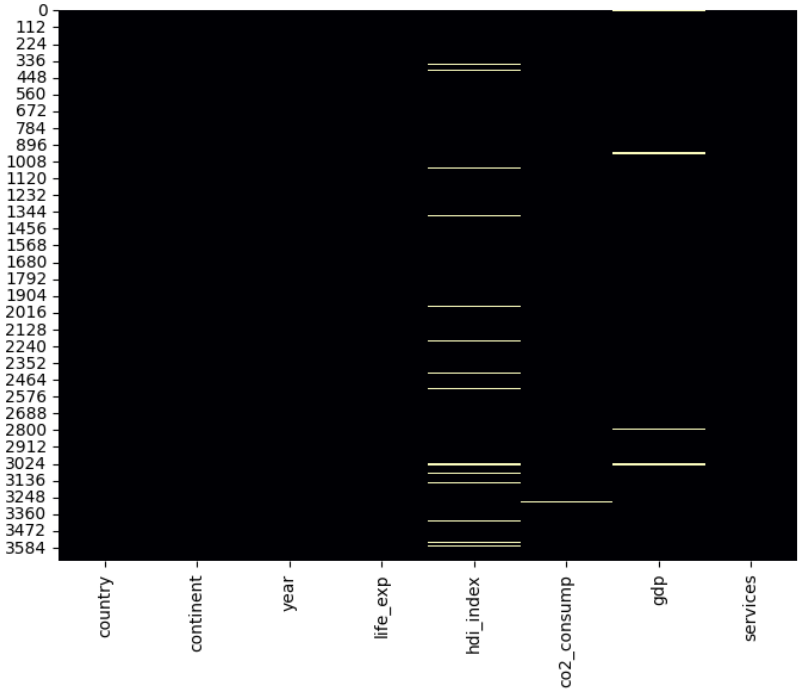


Figure 1 : Heatmap des valeurs manquantes

Concernant l'Indice de Développement Humain, les données manquantes concernent principalement les années 1998-1999 sur des pays de petite taille ou qui n'avaient pas de grande importance à l'époque pour récolter les données nécessaires, tels que St. Lucie, la Macédoine du nord, le Tchad ou encore la Guinée Equatoriale. D'autres pays comme le Soudan du Sud étaient

dans une situation politique et démographique compliquée ne permettant pas de connaitre l'indice de développement humain.

Pour les mêmes raisons, le Produit Intérieur Brut par habitant n'est pas indiqué pour certains pays. En ce qui concerne les données manquantes pour les émissions de CO2, on voit la présence du Timor-Leste (Timor Oriental) car, avant 2002, il s'agissait d'une province de l'Indonésie qui n'était donc pas encore indépendante.

Nous avons donc décidé de supprimer toutes les lignes ayant des valeurs manquantes. Nous sommes ainsi passés de 3 675 lignes à 3 532.

Analyse

Description des données

Nous avons commencé par faire une description des données en faisant la moyenne des données par pays, puis en regardant les plus hautes et les plus basses.

	Moyennes hautes	Moyennes basses
Espérance de vie	Japon (83)	République centrafricaine (46.7)
IDH	Norvège (0.934)	Nigéria (0.318)
Emission de CO2 (en tonnes)	Qatar (48.53)	République démocratique du Congo (0.03)
PIB par habitant (en $)	Luxembourg (95957)	Burundi (307)
% Secteur tertiaire	Hong Kong, Chine (84.85)	Burundi (8.15)

Tableau 1 : Moyennes hautes et basses pour chaque variable numérique

On peut remarquer que les moyennes hautes se situent toutes sur le continent européen ou asiatique, tandis que les moyennes basses proviennent toutes du continent africain. Cela semble logique étant donné que l'on retrouve les pays les plus développés dans les continents de l'hémisphère nord et les moins développés principalement sur le continent africain.

Après cela, nous avons affiché des histogrammes représentant la répartition des variables entre les continents. Pour cela, nous avons regroupé les données par pays puis par continent en calculant la moyenne pour chaque pays sur l'ensemble des années.

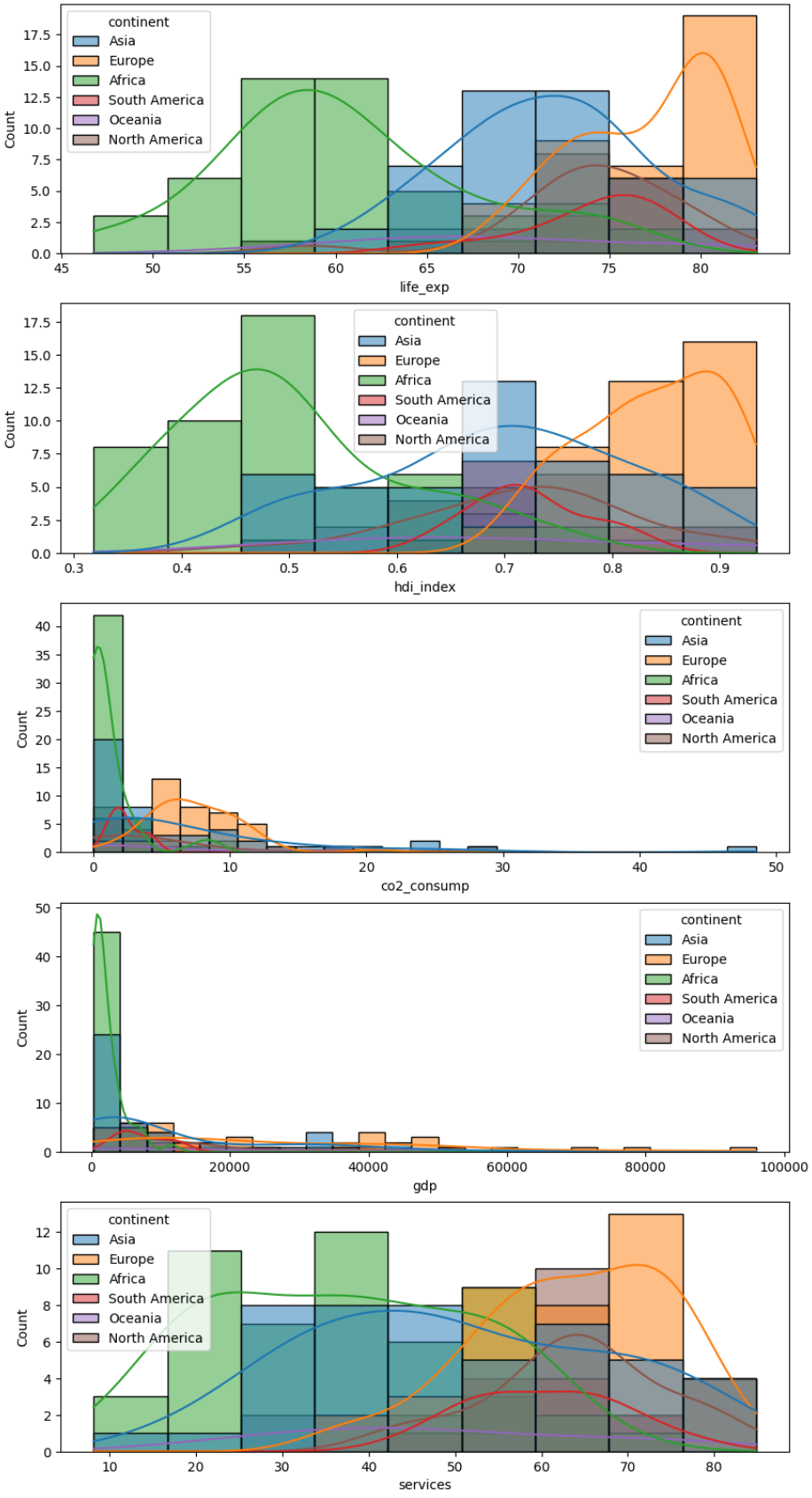


Figure 2 : Histogrammes des variables selon les continents

Nous pouvons voir, grâce à ces graphiques, que les pays africains sont situés du côté des moyennes basses, comme vu précédemment. Pour ce qui est de l'Asie, on peut remarquer de

grandes plages de valeurs pour chaque variable, ce qui indique que le continent contient aussi bien des pays développés que des pays en développement. Il s'agit de la même situation pour l'Amérique du Sud. L'Amérique du Nord pour sa part, est plus proche de l'Europe avec des plages de valeurs réduites et élevées. Cela indique que pour ces deux continents, les conditions de vie sont plutôt bonnes et qu'ils sont bien développés. Pour l'Océanie, il est difficile de savoir, car nous n'avons que peu de données en raison du peu de pays qui se situent sur ce continent.

D'un autre point de vue, si nous comparons les variables plutôt que les continents, on peut également remarquer qu'il existe peut-être une corrélation entre l'indice de développement humain et l'espérance de vie. Ces deux histogrammes sont très similaires sur la répartition des données et ont exactement la même échelle. On peut également y ajouter les services, bien que dans une moindre mesure.

Voici une autre représentation de ses résultats, avec cette fois l'utilisation de box-plots :

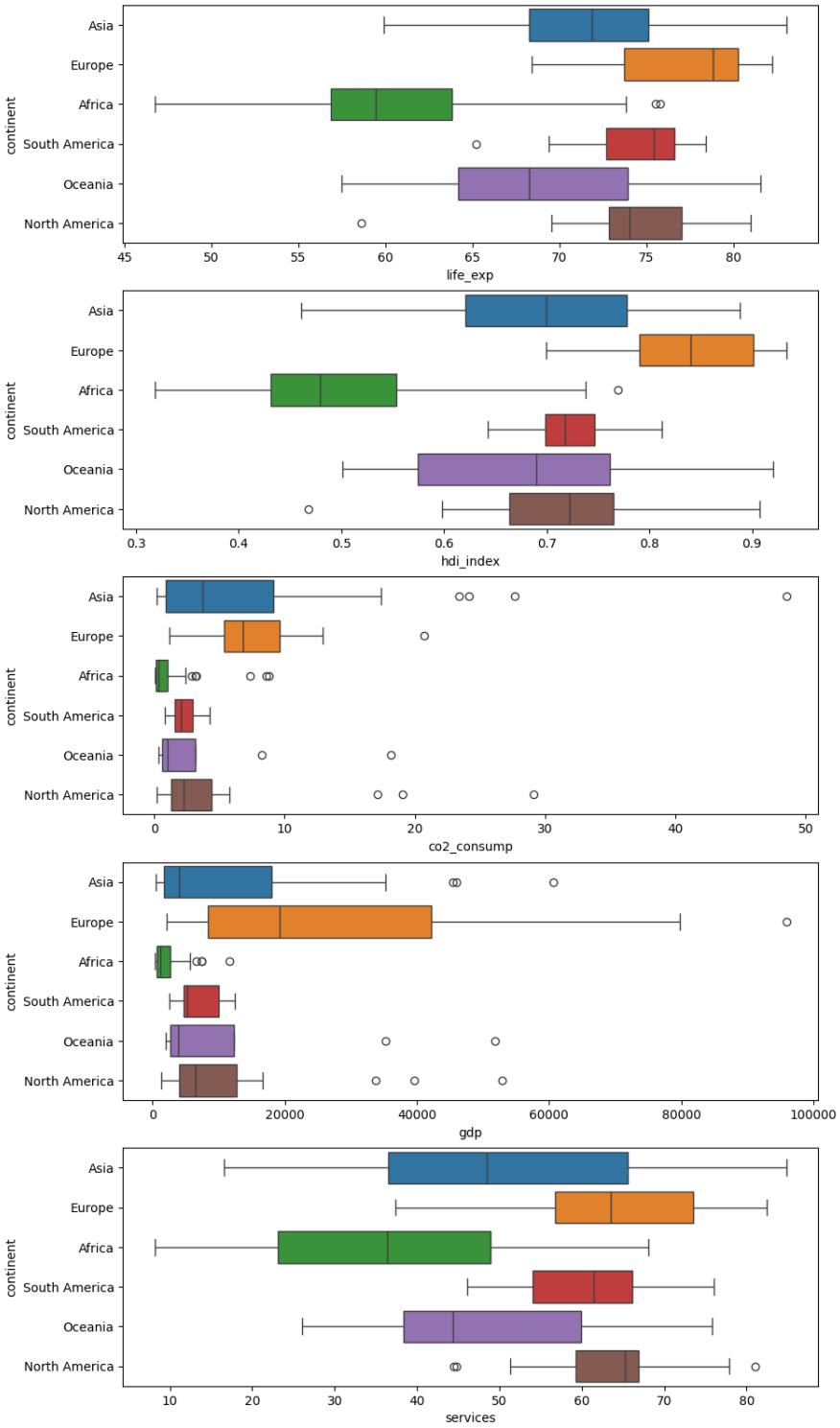


Figure 3 : Box-plot des variables selon les continents

Nous retrouvons les mêmes résultats, mais avec une représentation plus lisible que la précédente. Nous remarquons notamment qu'un certain nombre de pays se démarquent sur leur continent. Par exemple, en Europe, on observe que le Luxembourg possède une valeur pour le PIB très importante. Dans le même cas en Afrique, la Lybie a un indice de développement supérieur à la moyenne avec 0.769. A l'opposé, on constate qu'en Amérique du Sud, la plus petite espérance de vie est en Guyane avec 65.2 ans, ce qui est bien plus faible que la moyenne.

Pour revenir à la corrélation des variables, dont nous avons parlé précédemment, nous avons voulu afficher la matrice de corrélation.

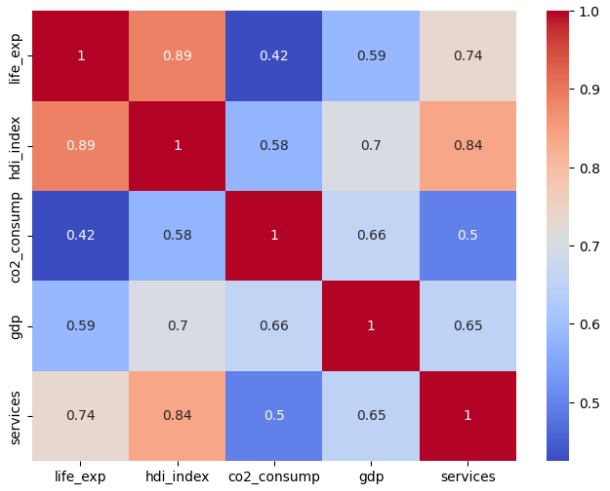


Figure 4 : Matrice de corrélation entre les variables numériques

Nous voyons bien qu'une forte corrélation existe entre l'IDH et l'espérance de vie. Ceci a du sens, car l'IDH est notamment calculé en prenant en compte l'espérance de vie. On observe également une corrélation importante entre l'IDH et le pourcentage de travailleurs dans le secteur tertiaire, qui peut être expliquée par le fait que les pays avec un IDH élevé ont souvent une économie tertiaire développée, avec le tourisme, par exemple. Ces corrélations sont aussi visibles si l'on décide de regarder l'organisation et la position des points sur des graphiques (Voir annexe 1).

On peut aussi essayer de regarder la corrélation entre les variables numériques et les différents continents pour voir si nous pouvons en tirer quelque chose.

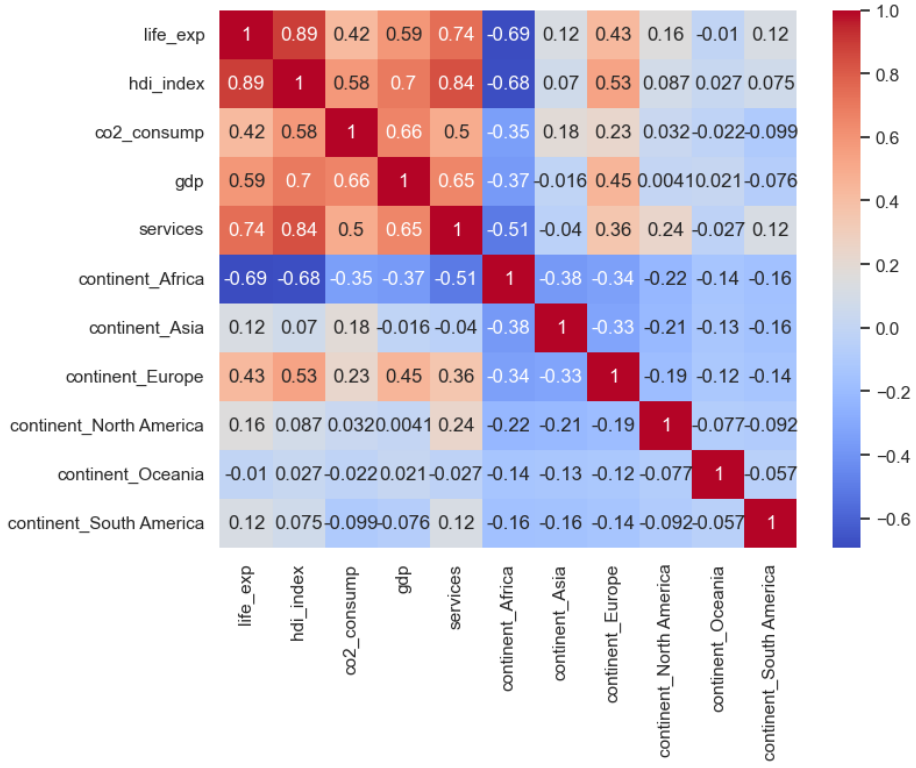


Figure 5 : Matrice de corrélation avec les continents

Nous pouvons nous apercevoir ici que les coefficients de corrélation entre nos variables et les continents sont très différents selon ces derniers. Ceux-ci vont de 0.23 à 0.53 pour l'Europe, tandis qu'ils sont situés entre -0.69 et -0.35 pour l'Afrique. Il est possible d'expliquer cela par le fait que l'Europe est un continent développé et possède donc des valeurs élevées pour l'IDH, le PIB par habitant ou l'espérance de vie, alors que l'Afrique est un continent en développement et possède donc des valeurs plus faibles pour ces mêmes variables. Cela rejoint donc ce que nous avons pu observer précédemment.

Clustering

Dans le but d'obtenir des informations supplémentaires sur les liens entre les pays et les continents, mais aussi de découvrir de potentielles anomalies, nous avons décidé de réaliser du clustering. Pour ce faire, nous avons commencé par normaliser nos données, car l'échelle n'est pas la même entre les variables, comme on peut le voir avec l'IDH et le PIB. De plus, nous allons utiliser trois méthodes différentes pour le clustering mais en ne prenant en compte que les colonnes numériques dans un premier temps.

Nous démarrons avec K-means en cherchant le nombre de clusters optimal. Après utilisation de la méthode du coude (voir **Erreur ! Source du renvoi introuvable.**), on détermine que 3 clusters sont intéressants pour le jeu de données. Nous obtenons les graphiques suivants qui montrent les dispositions de points avec l'affichage des clusters. Etant donné que ces graphiques ne sont pas visuellement simples à analyser, on choisit d'afficher également la répartition des clusters par continent.

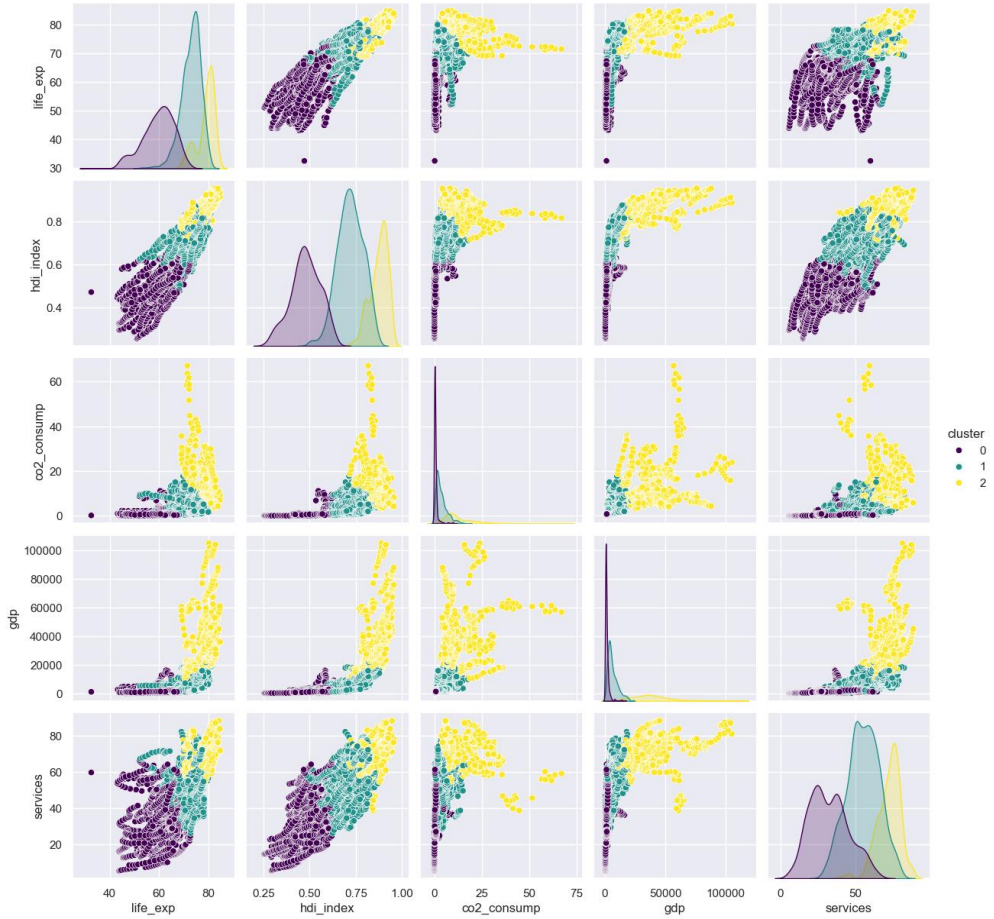


Figure 6 : Graphiques montrant les clusters avec K-Means

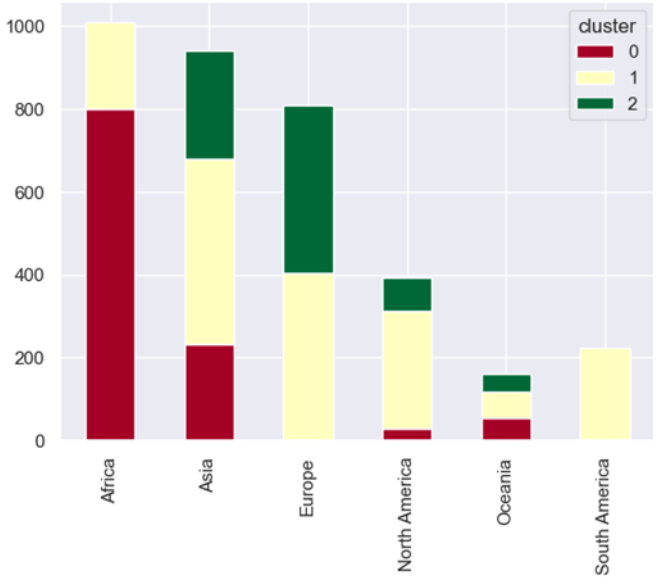


Figure 7 : Répartition des clusters pour chaque continent avec K-means

On remarque clairement que les clusters ne sont pas répartis de la même manière selon les continents. En effet, les pays africains sont majoritairement dans le cluster 0 alors qu'aucun d'entre eux ne se trouve dans le cluster 2. Au contraire, les pays européens sont exclusivement dans les clusters 1 et 2. On peut donc en déduire que ces clusters correspondent à des niveaux

de développement différents. Le premier correspond à des pays moins développés, tandis que les deux autres correspondent à des pays moyennement ou très développés.

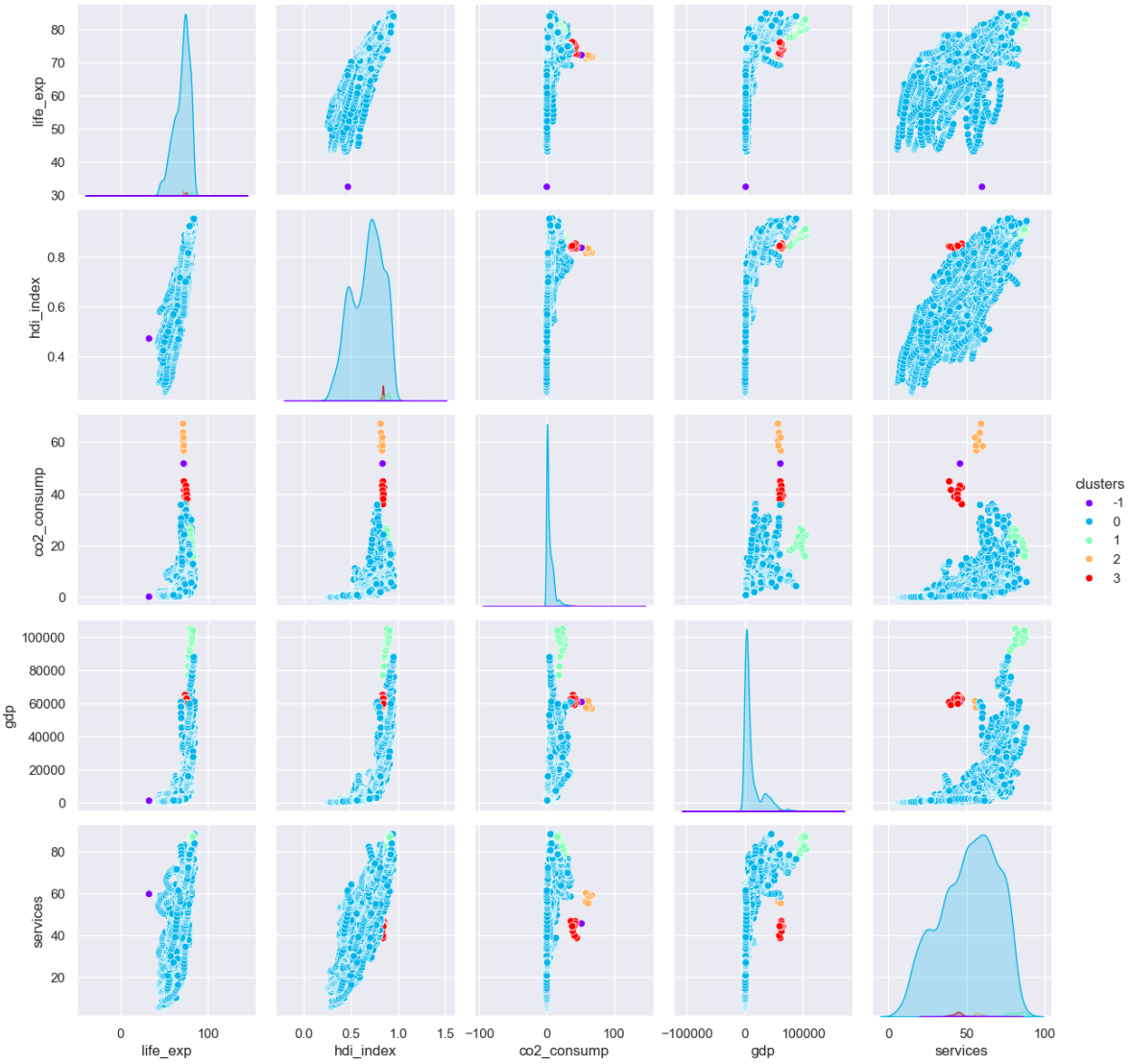


Figure 8 : Graphiques montrant les clusters avec K-Means

On essaye ensuite de faire un clustering avec DBScan pour voir les résultats.

Les résultats obtenus avec DBScan ne sont pas très concluants, car nous avons une grande majorité de points dans le même cluster et seuls les points qui sont très éloignés des autres sont dans un autre cluster. Néanmoins, lorsque l'on s'intéresse à ces points, on remarque plusieurs choses intéressantes.

Cluster	-1	0	1	2	3
Pays	Haïti (1) Qatar (1)	Reste des pays	Luxembourg (21)	Qatar (7)	Qatar (11)

Tableau 2 : Pays et leur nombre d'apparitions pour chaque cluster

Nous pouvons tout d'abord constater la présence d'Haïti dans un cluster à part. Lorsque l'on regarde les valeurs de ses colonnes, on observe qu'en 2010, l'espérance de vie descend à 32.5 ans avant de remonter l'année suivante. En effectuant quelques recherches, on constate que

cette baisse est due à un tremblement de terre qui a eu lieu cette année-là. Il a causé la mort de plusieurs centaines de milliers de personnes et donc une baisse de l'espérance de vie.

On peut également s'intéresser à la présence du Luxembourg dans un cluster à part. En effet, le pays est présent dans le cluster 1 pour chacune des années de notre jeu de données. Lorsqu'on regarde les valeurs de ses colonnes, on observe que le pays a des valeurs plus élevées que tous les autres pays pour le PIB par habitant ainsi que pour l'émission de CO2, ce qui est bien le cas dans la réalité. On peut donc en déduire que le clustering a bien fonctionné pour ce pays.

Enfin, on voit que le Qatar se retrouve dans plusieurs clusters, mais séparé des autres pays. Etant donné que c'est un pays qui émet beaucoup de CO2, c'est pour cette raison que les points le représentant sont éloignés des autres et se retrouvent dans des clusters à part.

Nous avons terminé par tester une autre méthode de clustering avec le modèle de mélanges gaussien avec deux composantes, car il s'agissait du meilleur score. On imagine que les pays sont répartis en deux clusters en fonction de leur niveau de développement, mais cela ne nous apprend rien de plus par rapport à K-means.

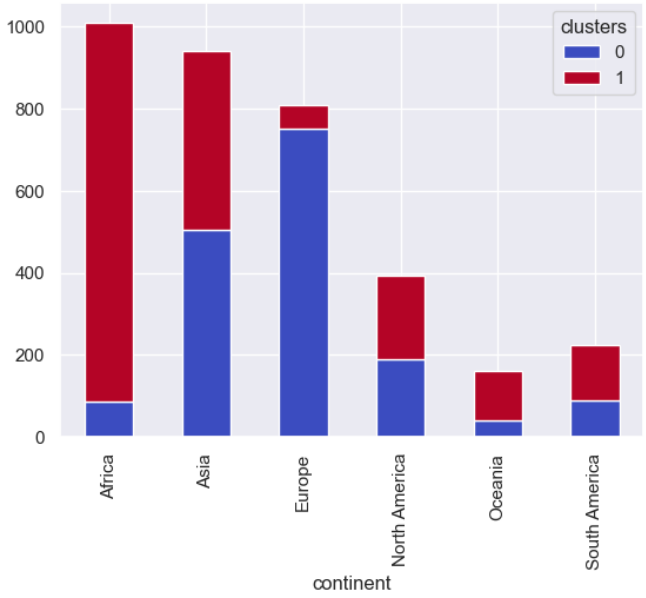


Figure 9 : Répartition des clusters pour chaque continent avec le GMM

Transformation des données

Encodage des variables catégorielles

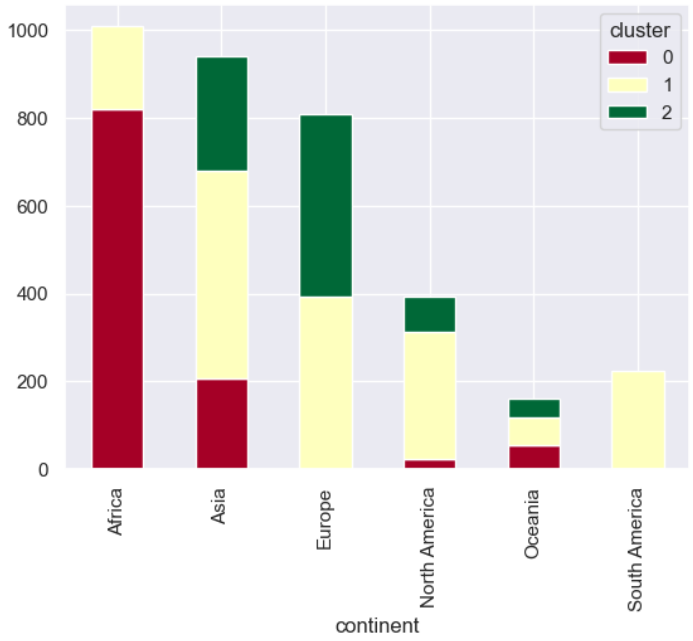


Figure 10 : Répartition des clusters pour chaque continent avec toutes les colonnes

On décide de garder K-means comme méthode de clustering car c'est celle qui nous a donné les meilleurs résultats. On commence par utiliser un objet ColumnTransformer et utiliser OneHotEncoder pour transformer les données et prendre en compte cette fois toutes les colonnes pour trouver le nombre optimal de clusters.

A part pour quelques points qui changent de clusters, nous n'observons pas de grande différence que ce soit en prenant en compte toutes les colonnes ou seulement les colonnes numériques.

Réduction de dimensions

Dans cette partie, nous essayons de réduire la dimension de nos données pour voir si on peut obtenir de meilleurs résultats. Nous utilisons d'abord PCA puis t-SNE.

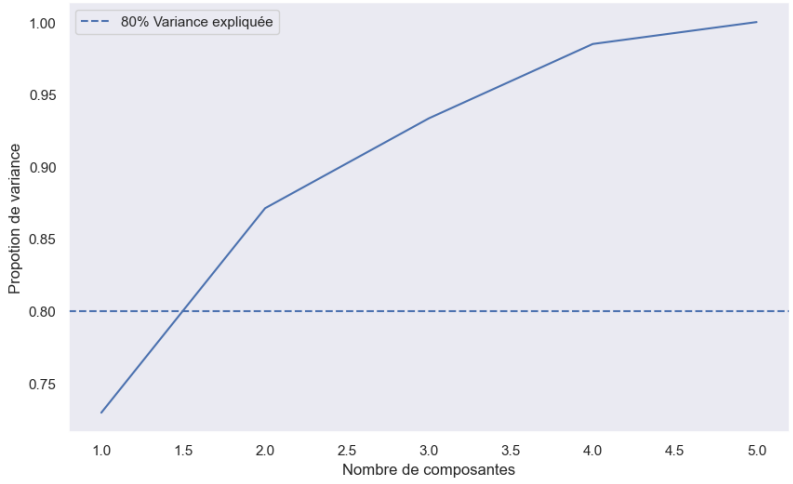


Figure 11 : Courbe pour trouver la variance

En fixant un seuil à 80%, on trouve que l'on peut garder seulement deux composantes principales pour garder suffisamment d'informations. On obtient donc ces graphiques à deux dimensions :

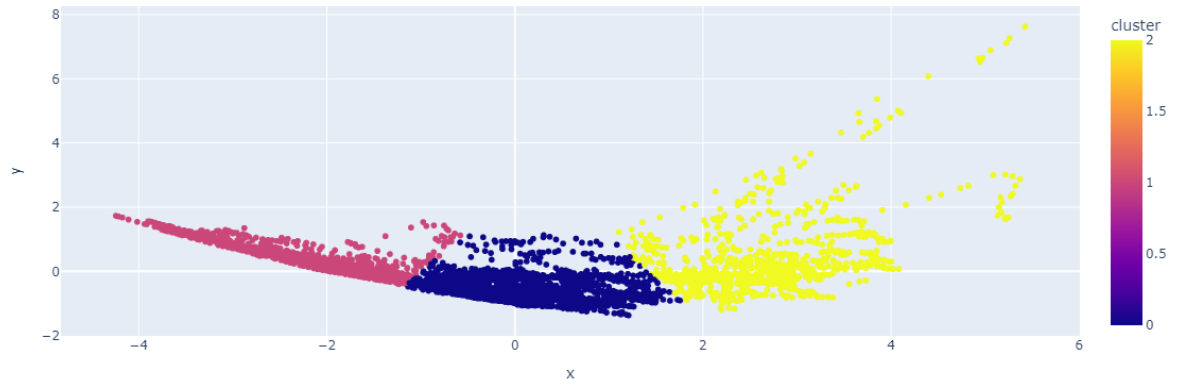


Figure 12 : Graphique obtenu en passant à 2 dimensions avec PCA

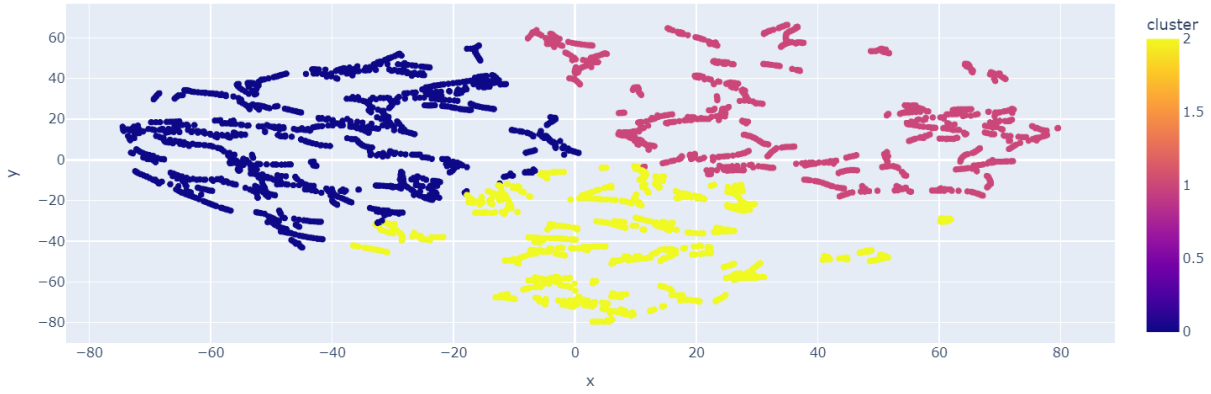


Figure 13 : Graphique obtenu en passant à 2 dimensions avec t-SNE

Cela nous montre une meilleure visualisation des clusters obtenus. On retrouve les mêmes clusters que précédemment, mais on peut mieux les distinguer.

Analyse temporelle

Pour conclure, nous avons voulu faire un peu d'analyse temporelle sur nos variables. Pour cela, nous avons commencé par voir l'évolution de nos variables avec les plus fortes corrélations dans le temps. Pour rappel, il s'agit de l'IDH avec les services ainsi que l'IDH avec l'espérance de vie :

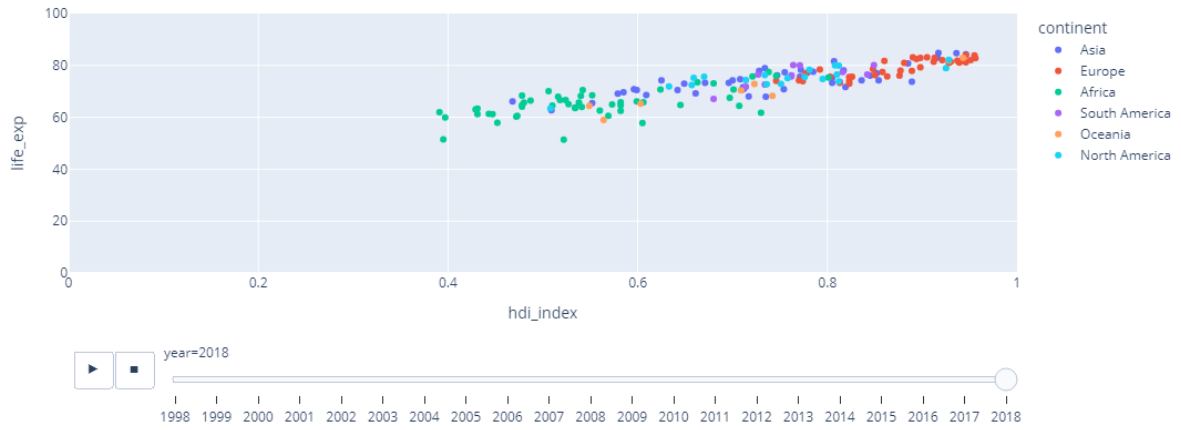


Figure 14 : Evolution de l'espérance de vie corrélée à l'IDH dans le temps

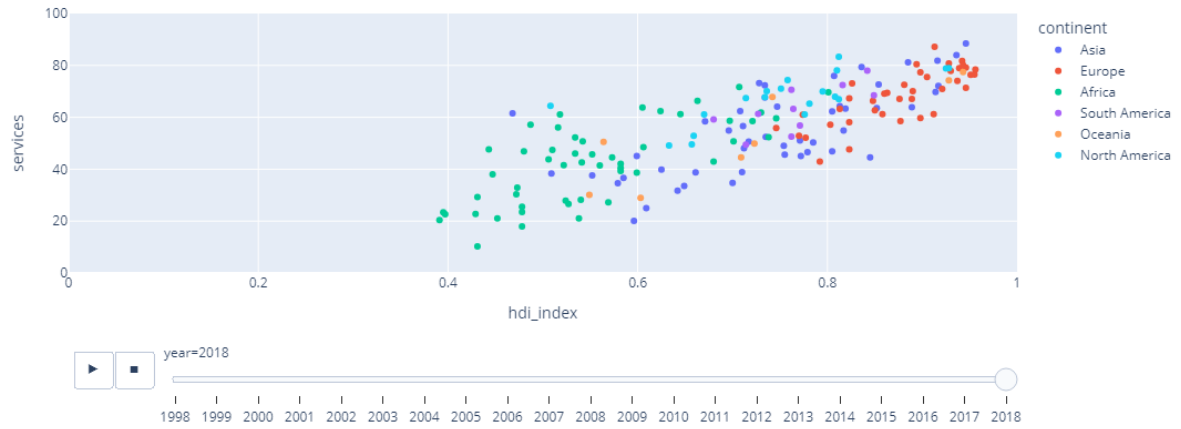


Figure 15 : Evolution des services corrélés à l'IDH dans le temps

L'animation complète est disponible dans notre code. Avec ces deux affichages, nous pouvons voir que les variables impliquées évoluent ensemble au fil du temps. Ce qui prouve bien une forte corrélation.

Par la suite, nous avons regardé les moyennes par année et continent.

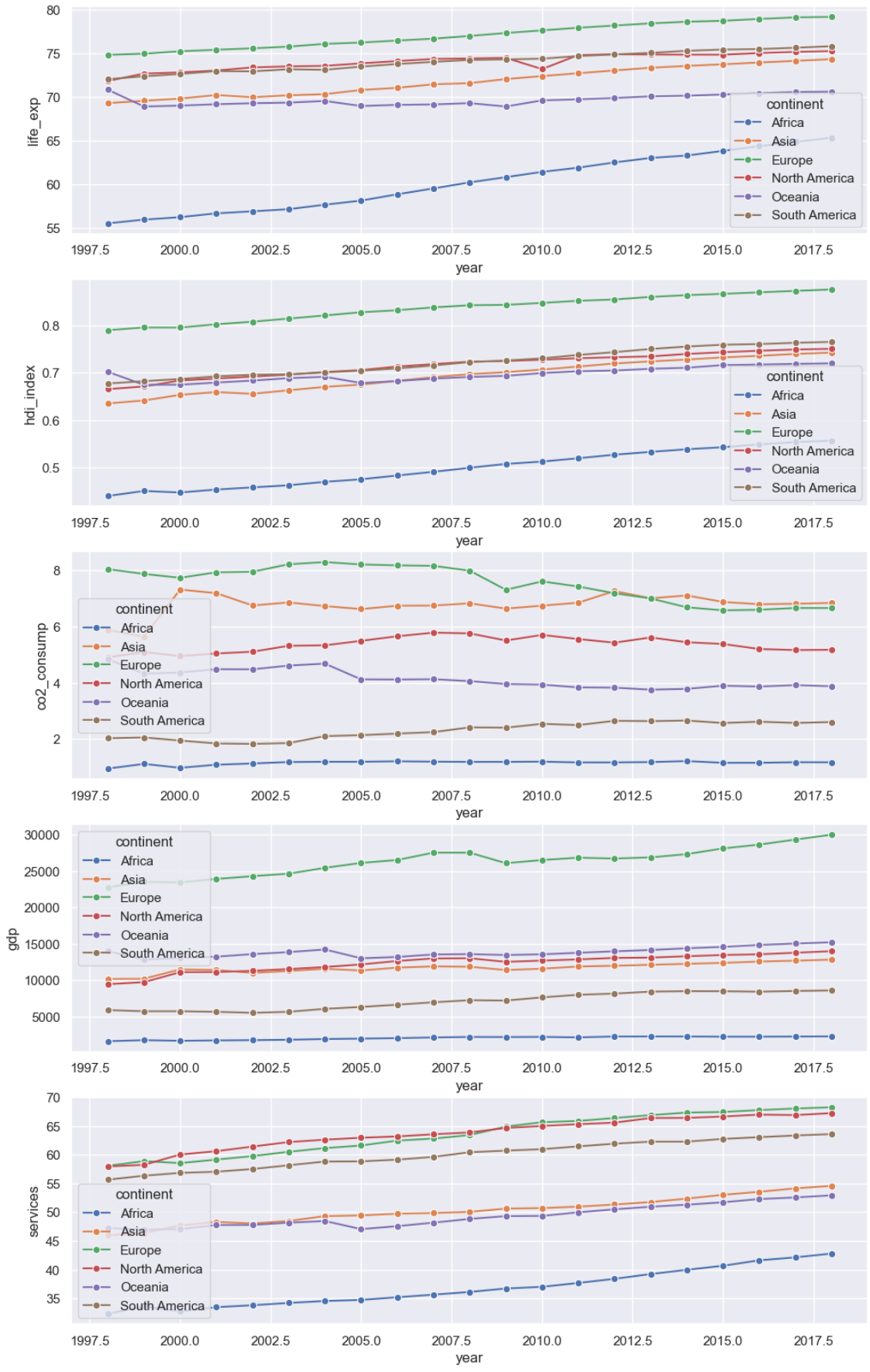


Figure 16 : Lineplot des moyennes par années selon les continents

Grâce à ces graphiques, on peut remarquer que l'espérance de vie ne fait qu'augmenter, bien que nous ayons un pic décroissant en 2010 pour l'Amérique du Nord. Cela correspond au séisme à Haïti dont nous avons parlé précédemment. L'augmentation globale est logique, au fil du temps la médecine s'améliore et de nouvelles découvertes sont faites. Cela permet d'améliorer les conditions de vie dans le monde et ainsi les personnes vivent plus longtemps.

Concernant les émissions de CO2, à part pour l'Asie, tous les continents sont en baisse ou restent stables dans leurs émissions. Pour ces derniers, cela semble logique étant donné les efforts ou les contraintes imposées à chacun pour réguler les émissions de gaz à effet de serre et limiter le réchauffement climatique. Pour l'Asie, qui est cependant l'un des principaux émetteurs de CO2, la moyenne ne fait que croitre. Cela peut être dû au fait qu'une grande partie des usines de production de masse sont localisées là-bas et émettent beaucoup, en plus des émissions qui proviennent du transport des marchandises produites.

Pour le produit intérieur brut et le pourcentage de travailleurs dans les secteurs tertiaires, il y a une augmentation globale ou une stabilité dans les moyennes de chaque continent.

Nous avons voulu, finalement, faire une analyse plus fine en faisant de la détection d'anomalies dans le but de voir si des pays avaient des valeurs incohérentes, sans analyse approfondie, sur certaines années. Pour cela, nous avons mis en place la méthode de la détection d'anomalie par moyenne glissante. Cette méthode permet de trouver tous les résultats anormaux compris entre la valeur actuelle et plus ou moins la fenêtre de temps. En plus de cela, il faut rajouter un seuil pour ne pas prendre en compte le moindre petit écart de valeur.

Nous avons fait cette détection pour trois colonnes du dataset : l'espérance de vie, l'émission de CO2 et l'indice de développement humain.

Voici nos résultats pour la détection d'anomalies sur l'espérance de vie :

country		country	year	life_exp	moving_avg	deviation
Burundi	515	Burundi	2006	54.5	52.425	2.075
Haiti	1366	Haiti	2009	60.7	53.350	7.350
	1367	Haiti	2010	32.5	53.350	-20.850
	1368	Haiti	2011	60.0	53.650	6.350
	1369	Haiti	2012	61.4	53.875	7.525
Liberia	1845	Liberia	2014	59.9	62.175	-2.275
Libya	1862	Libya	2011	73.2	75.400	-2.200
Myanmar	2206	Myanmar	2008	57.7	61.575	-3.875
Palestine	2445	Palestine	2010	73.7	71.625	2.075
Sri Lanka	2964	Sri Lanka	2004	69.0	72.275	-3.275

Figure 17 : Tableau des anomalies pour l'espérance de vie

Comme dit précédemment, Haïti a connu un très gros tremblement de terre en 2010. Nous pouvons voir les conséquences sur la prédiction de l'espérance de vie pour les années suivantes ici. En 2005, le Burundi a mis fin à une guerre civile qui frappait le pays depuis 1993. Cela explique que l'année suivante l'espérance de vie ait augmenté. Pour le Liberia, 2014 correspond à

l'épidémie d'Ebola qui touchait l'Afrique de l'Ouest. La Lybie a connu une intervention militaire de la part de l'ONU en 2011, ce qui a pu entraîner des conséquences négatives sur l'espérance de vie. La Birmanie (Myanmar) a connu en 2008 un cyclone qui a fait des dizaines de milliers de morts et disparus. En ce qui concerne la Palestine, nous n'avons pas pu trouver d'évènement marquant qui aurait pu faire croitre l'espérance de vie. Enfin, en 2004, le Sri Lanka a été touché par un séisme de magnitude 9.1, suivi d'un tsunami qui a fait 31 000 morts dans le pays. Tous ces évènements ont un impact direct sur la fluctuation de l'espérance de vie, que ce soit à cause de catastrophes naturelles ou de conflits.

Passons maintenant à l'analyse pour l'émission de CO2. Nous avons mis un seuil à 3 :

country		country	year	co2_consump	moving_avg	deviation
Bahrain	206	Bahrain	2001	20.2	24.600	-4.400
Brunei	455	Brunei	2007	22.2	18.000	4.200
	456	Brunei	2008	23.7	19.725	3.975
	461	Brunei	2013	18.9	22.200	-3.300
Mongolia	2135	Mongolia	2013	15.1	11.335	3.765
Qatar	2605	Qatar	2004	56.7	59.875	-3.175
	2607	Qatar	2006	61.8	57.200	4.600
	2609	Qatar	2008	44.8	49.950	-5.150
	2613	Qatar	2012	42.5	39.125	3.375
	2614	Qatar	2013	36.0	40.200	-4.200
Singapore	2817	Singapore	2009	11.8	8.780	3.020
United Arab Emirates	3333	United Arab Emirates	2000	35.7	30.275	5.425
	3335	United Arab Emirates	2002	24.1	29.700	-5.600

Figure 18 : Tableau des anomalies pour l'émission de co2

Il est compliqué de statuer sur d'où viennent ses écarts dans l'émission de dioxyde de carbone. Cela peut venir d'évènements qui sont survenus dans le pays, d'une déviation tout à fait normale ou de tout autre chose. Nous ne pouvons pas conclure sur la provenance de ces résultats.

Enfin, en ce qui concerne l'IDH, avec un seuil de 0.02, nous avons trouvé trois valeurs :

country		country	year	hdi_index	moving_avg	deviation
Libya	1862	Libya	2011	0.764	0.78750	-0.02350
	1866	Libya	2015	0.697	0.71825	-0.02125
Timor-Leste	3164	Timor-Leste	2008	0.616	0.59375	0.02225

Figure 19 : Tableau des anomalies pour l'IDH

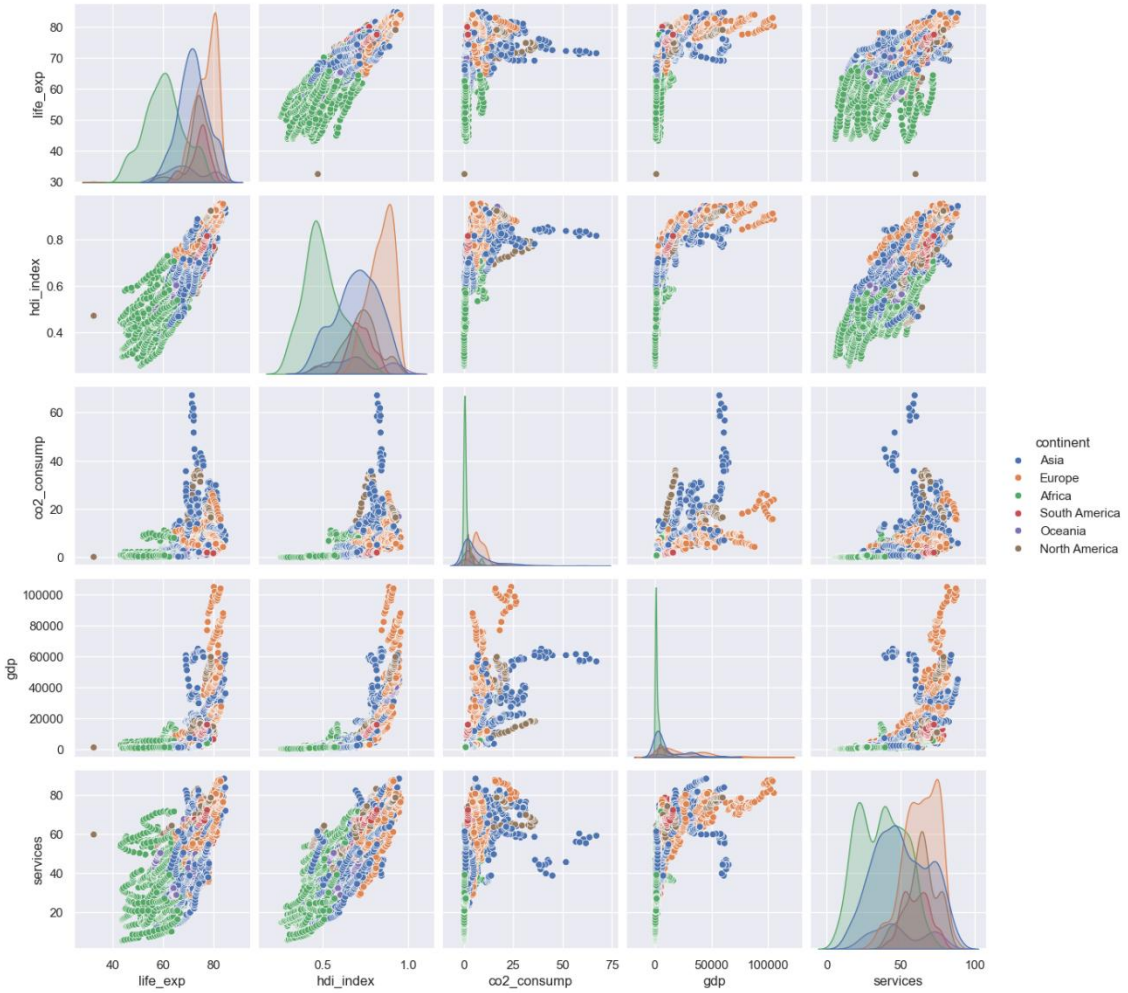
Pour la Libye, comme dit précédemment pour l'espérance de vie, 2011 correspond à l'intervention militaire organisée par l'ONU. La seconde valeur, en 2015, peut avoir été impactée par la deuxième guerre civile libyenne. Ce genre de conflit peut avoir des impacts directs sur l'IDH d'un pays. Pour le Timor Oriental (Timor-Leste), nous n'avons pas trouvé d'explication à cette hausse.

Annexe

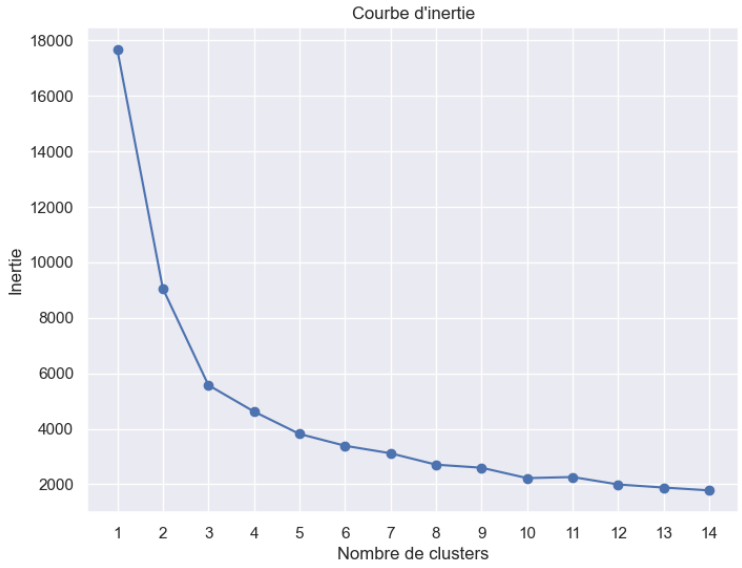
Répartition

	Thomas	Angéline
Traitement des données		
Description		
Clustering		
Transformation		
Temporelle		
Rapport		

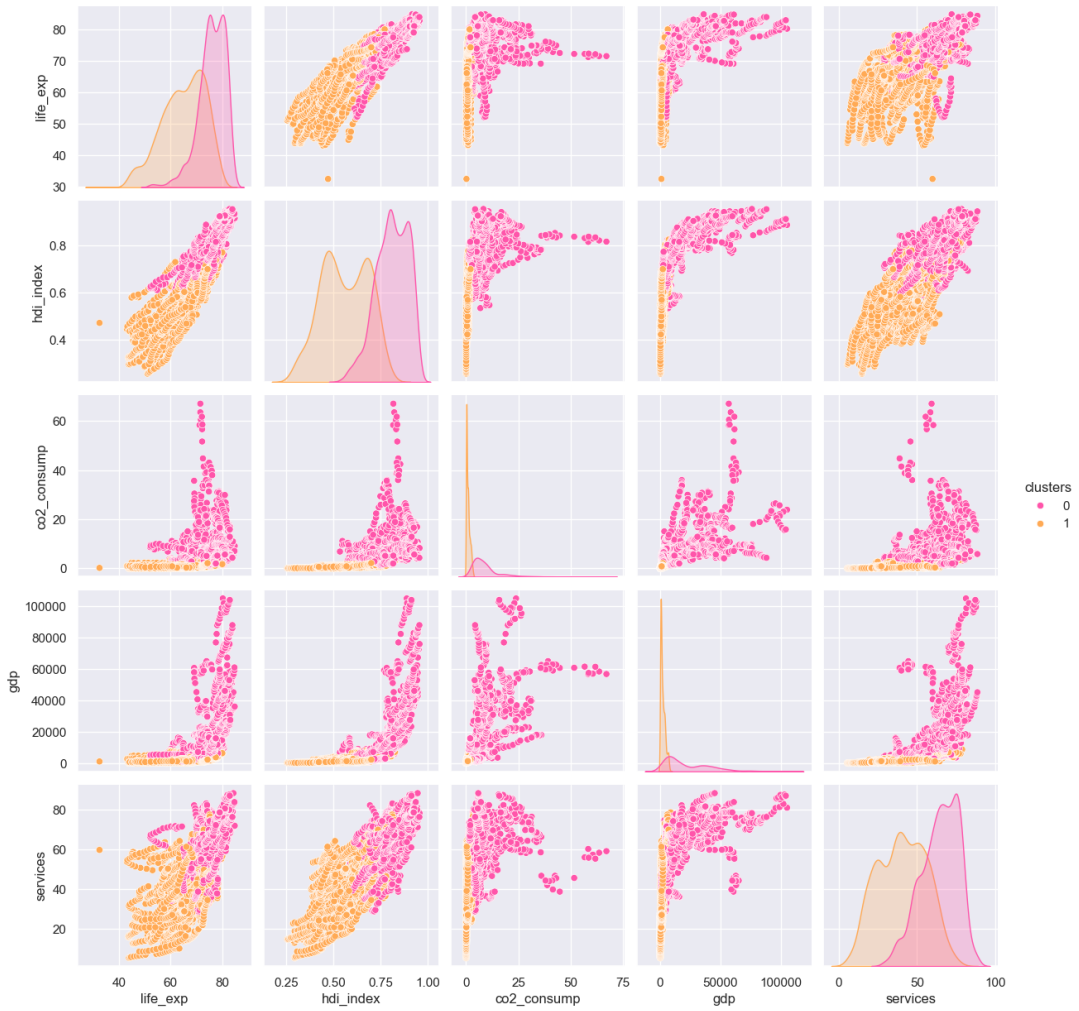
Images



Annexe 1 : Graphiques de corrélation entre les variables numériques



Annexe 2 : Méthode du coude pour obtenir le nombre de cluster optimal



Annexe 3 : Graphiques pour le clustering avec le mélange gaussien