



DataScientest • com

Rapport Technique d'évaluation

PROJET WORLD HAPPINESS

Promotion Data Analyst BootCamp

AOÛT – OCTOBRE 2021

Participants :

Eleonora FABRIS

Loris SEDEAUD

Pierre THOMAS

Mentor :

Rania KHOUADJA

Table des matières

Introduction	3
Contexte :	3
Objectifs du projet :	3
Expertise du groupe sur le sujet et la problématique adressée :	4
Preprocessing	5
Datasets :	5
Notes explicatives :	7
Dataviz	8
Modélisation	21
Conclusion	23
Annexes	24
Diagramme de GANTT :	24

Introduction

Contexte :

Le sujet de notre projet porte sur l'analyse des données collectées par le World Happiness Report. Cette enquête, publiée chaque année, a pour objectif d'estimer le bonheur des pays du monde à l'aide de mesures socio-économiques telles que le PIB par habitant, l'espérance de vie, la liberté de choix, la générosité ou encore la perception de la corruption.

Le "Ladder score" représente ici le score global de bonheur par pays et s'obtient par la somme des différentes variables "Explained by" à laquelle s'additionne la variable "Dystopia + Residual".

Nous allons durant ce projet, déterminer les combinaisons de facteurs qui permettent d'expliquer pourquoi certains pays, voire certaines régions du monde, sont mieux classé(e)s que d'autres.

Objectif du projet :

L'objectif du projet est de comprendre quels sont les facteurs impactant sur le score de bonheur et de pouvoir prévoir celui-ci à partir des données brutes de chaque pays.

Également, via différents modèles prédictifs testés, nous allons vérifier si ces mêmes modèles peuvent être appliqués à l'ensemble des pays ou si ils ne sont fonctionnels que pour certains d'entre eux. Nous déterminerons les combinaisons de facteurs qui permettent d'expliquer pourquoi certains pays sont mieux classés que d'autres.

Expertise du groupe sur le sujet et la problématique adressée :

Loris : issu du marketing et des achats principalement dans l'univers des biens de grande consommation en tant que chef de produit, cette approche d'analyse de données est tout à fait nouvelle pour moi. J'ai pris plaisir à mettre en pratique les compétences récemment acquises durant cette formation pour appliquer les différentes étapes d'un projet d'étude comme celui-ci, sur le thème du bonheur.

Pierre : formateur en gestion de projet, la réalisation de notre projet était intéressante pour découvrir une méthode plus approfondie que ce que je connaissais dans l'analyse des données et des relations entre les différentes variables.

Eleonora : Data Manager pendant plus de 5 ans dans l'aérospatial, ce projet m'a permis de travailler sur les données d'un nouveau secteur et de mettre en pratique les connaissances acquises tout le long de ce parcours de Data Analyst notamment en modélisation et webscraping.

Preprocessing

Data :

Les datasets utilisés proviennent du site Kaggle, à l'adresse suivante :
<https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021>

Le premier jeu de données est le fichier “world-happiness-report-2021” qui contient l'étude du bonheur pour l'année 2021 (l'étude la plus récente).

Ce jeu de données contient au total 149 lignes.

Les variables sont les suivantes :

- Country name
- Regional indicator
- Ladder score
- Standard error of ladder score
- upperwhisker
- lowerwhisker
- Logged GDP per capita
- Social support
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption
- Ladder score in Dystopia
- Explained by: Log GDP per capita
- Explained by: Social support
- Explained by: Healthy life expectancy
- Explained by: Freedom to make life choices
- Explained by: Generosity
- Explained by: Perceptions of corruption
- Dystopia + residual

Le second jeu de données est le fichier “world-happiness-report” qui, lui, contient l'étude du bonheur de l'année 2005 à 2020.

Il contient au total 1949 lignes.

Les variables sont les suivantes :

- Country name
- year

- Life Ladder
- Log GDP per capita
- Social support
- Healthy life expectancy at birth
- Freedom to make life choices
- Generosity
- Perceptions of corruption
- Positive affect
- Negative affect

Ces données sont disponibles librement via le site kaggle. Les deux dataset sont de taille relativement petite. Concernant le jeu de données de 2021, il n'y a eu aucun data cleaning d'effectué, le dataset ne contenant ni erreur ni valeur aberrante. Concernant le second jeu de données de 2005 à 2020, il arrive que pour certaines années, certaines valeurs soient vides. Ces valeurs manquantes ont été remplacées par la moyenne de la variable du pays concerné.

Pour le jeu de données 2021, le ladder score (score de bonheur) s'obtient par la somme des variables "Explained by" + "Dystopia + residual". Nous avons alors fait le choix de ne garder que ces variables et avons décidé de supprimer les autres.

Notre variable cible est donc la variable "Ladder score". Pour avoir un ordre d'idée, pour l'année 2021, le score de bonheur le plus élevé est 7.84 et le moins élevé est 2.52 pour une moyenne de 5.53 parmi les 149 lignes.

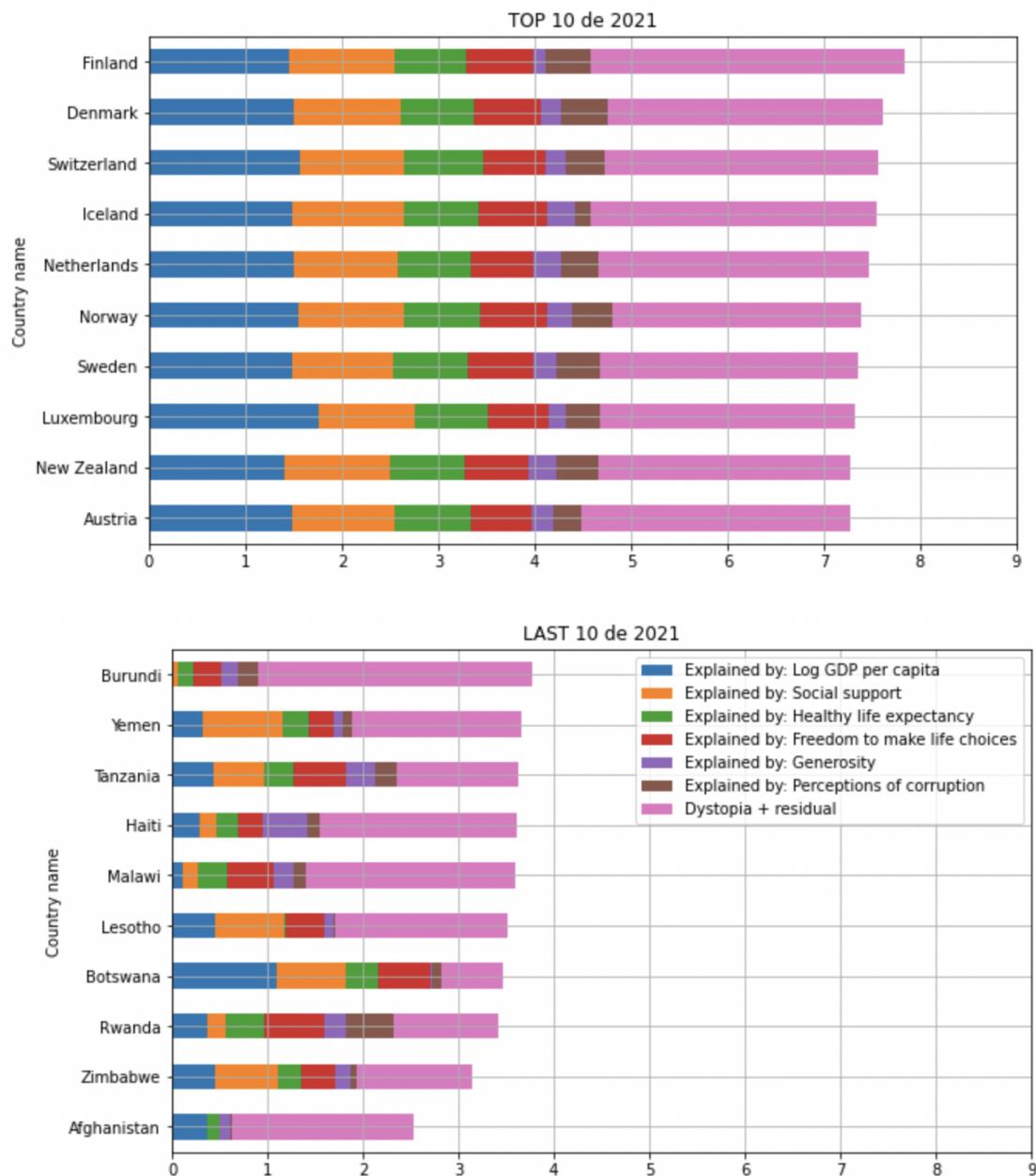
Il existe des relations entre les différentes variables. Premièrement, entre la variable cible et les variables explicatives, la variable cible étant la somme, comme indiqué précédemment, de certaines d'entre elles. Il existe également des relations entre variables explicatives car certaines fonctionnent comme "paires" : par exemple, la variable "Social support" et la variable "Explained by: Social support". En dehors de ces dépendances, il existe des liens logiques entre variables; en effet, lorsque le PIB/habitant est élevé, les autres variables explicatives ont plus de chances d'être élevées également.

Notes explicatives :

- Le Ladder score (score de bonheur) s'obtient par la somme des variables “Explained by” + “Dystopia + residual”.
- La variable “Social Support” est la moyenne nationale à des réponses binaires (0 ou 1) à la question : “Si vous aviez des problèmes, avez-vous de la famille ou des amis sur lesquels compter en cas de besoin ?”.
- La variable “Freedom to make life choices” est la moyenne nationale des réponses binaires à la question : “Êtes-vous satisfait de votre liberté de choisir ce que vous faites de votre vie ?”.
- La variable “Generosity” est le résidu de la régression de la moyenne nationale des réponses à la question “Avez-vous fait un don à un organisme de charité au cours du mois dernier ?” sur le PIB/habitant.
- La variable “Perceptions of corruption” est la moyenne des réponses binaires à deux questions : “La corruption vous paraît-elle répandue au sein du gouvernement ?” et “L'est-elle au sein des entreprises ?”.
- La variable “Ladder score in Dystopia” représente le score d'un pays versus Dystopia (\neq Utopia). Il s'agit-là d'un pays imaginaire contenant les personnes les moins heureuses du monde, le but étant d'avoir une référence par rapport à laquelle tous les pays peuvent être comparés favorablement (aucun pays n'a de plus mauvais scores).
- La variable “Dystopia + residual” correspond aux résidus ou composantes inexpliquées qui diffèrent pour chaque pays, reflétant quand quelle mesure les six variables sur-expliquent ou sous-expliquent les évaluations de la vie moyenne.
- Les variables “Positive affect” et “Negative affect” représentent respectivement la fréquence moyenne de bonheur, de rire et de plaisir d'un côté, la fréquence moyenne d'inquiétude, de tristesse et de colère de l'autre.

Dataviz

1/ Top 10 & Last 10



Observations :

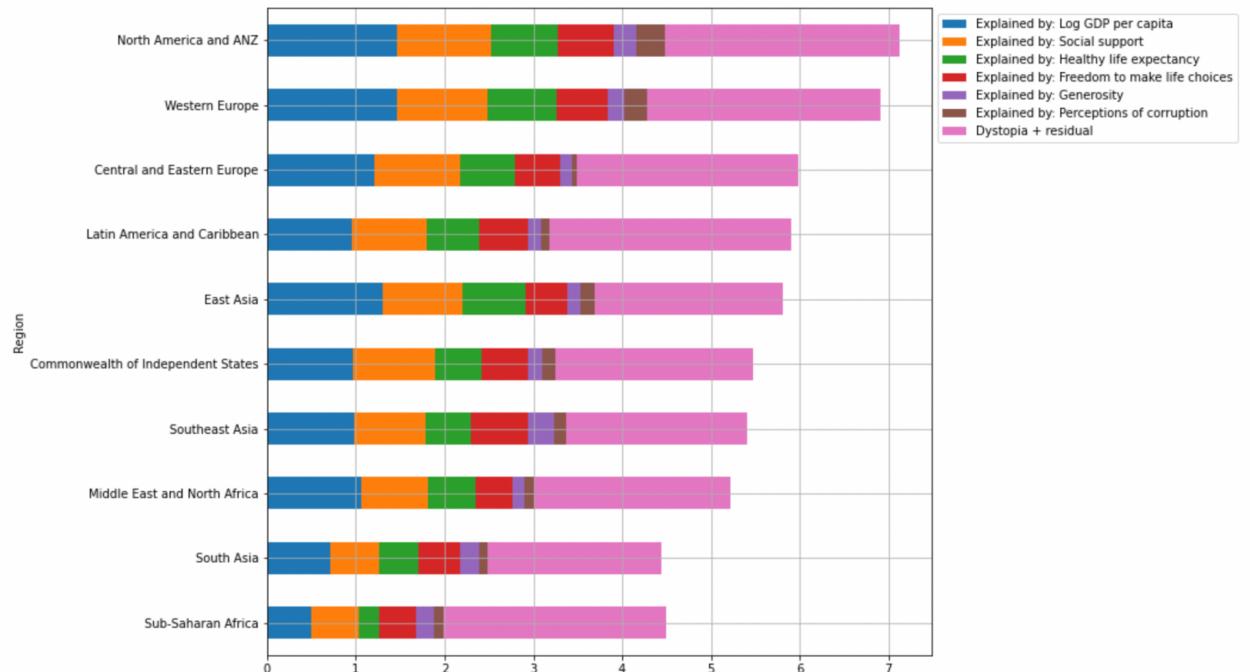
Nous avons souhaité avoir une idée précise du poids de chaque variable qui compose le score de bonheur parmi les 10 premiers et 10 derniers pays du classement via deux barplot empilés.

Mise à part la variable "Dystopia + residual" qui additionne la comparaison du pays versus Dystopia (\neq Utopia, un pays imaginaire où le bonheur est nul) et les composantes

inexpliquées, nous pouvons apercevoir, pour les 10 premiers pays, que le critère le plus important pour “être heureux” est le PIB/habitant, suivi par l'aide sociale, l'espérance de vie et la liberté de choix. Enfin, dans une moindre mesure, interviennent la générosité et la perception de la corruption.

Pour les 10 pays les moins bien classés, la variable PIB/habitant est la plus importante pour seulement deux d'entre eux. L'aide sociale est la plus importante pour trois pays et la liberté de choix pour quatre, ce qui marque une rupture de vision du bonheur entre habitants de pays au Ladder score élevé et bas.

2/ Classement par région



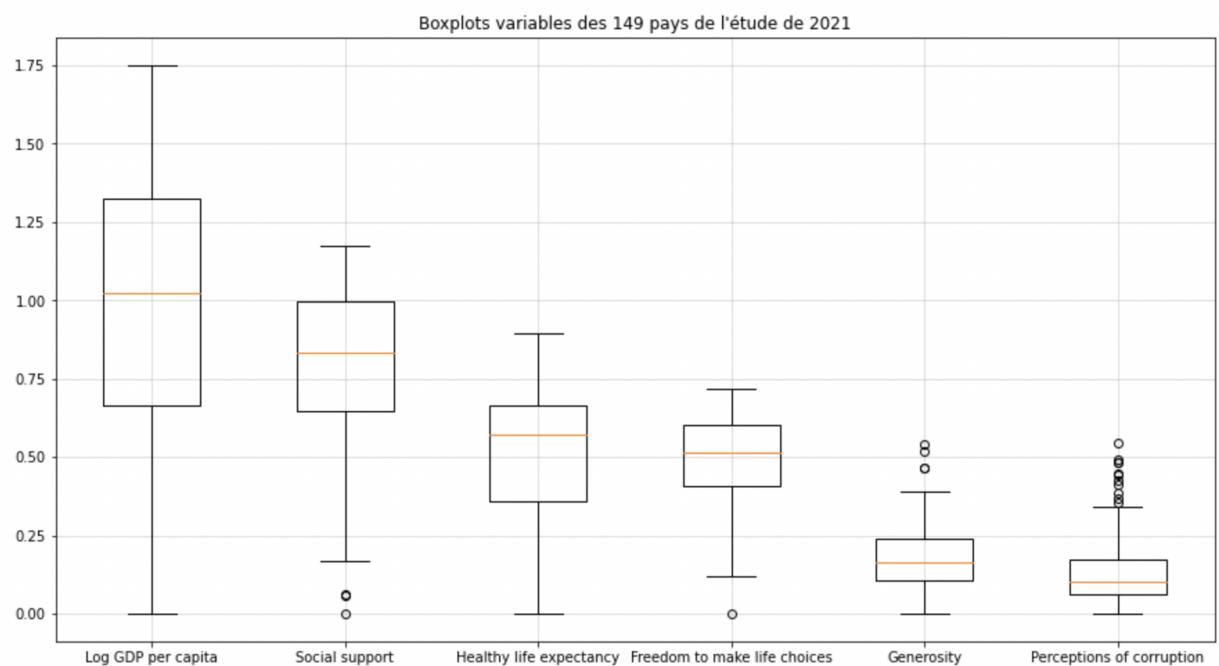
Observations :

Ici, nous souhaitions connaître le classement des régions du monde les plus heureuses ainsi que le poids des variables composant le Ladder score grâce à la moyenne des notes des pays appartenant à chacune d'elles.

Les trois régions du monde où le Ladder score est le plus élevé sont l'Amérique du nord + Australie + Nouvelle-Zélande, l'Europe de l'ouest puis l'Europe centrale et du sud. A l'inverse, celles où il est le plus bas sont le Moyen-Orient + Afrique du nord, l'Asie du sud et enfin l'Afrique Subsaharienne.

Pour les régions du monde détenant un Ladder score élevé, le PIB/habitant est la variable la plus élevée, suivie par l'aide sociale et l'espérance de vie. Pour celles ayant un Ladder score bas, le PIB/habitant reste la variable majeure mais dans une bien moindre mesure, souvent à égalité avec l'aide sociale et la liberté de choix.

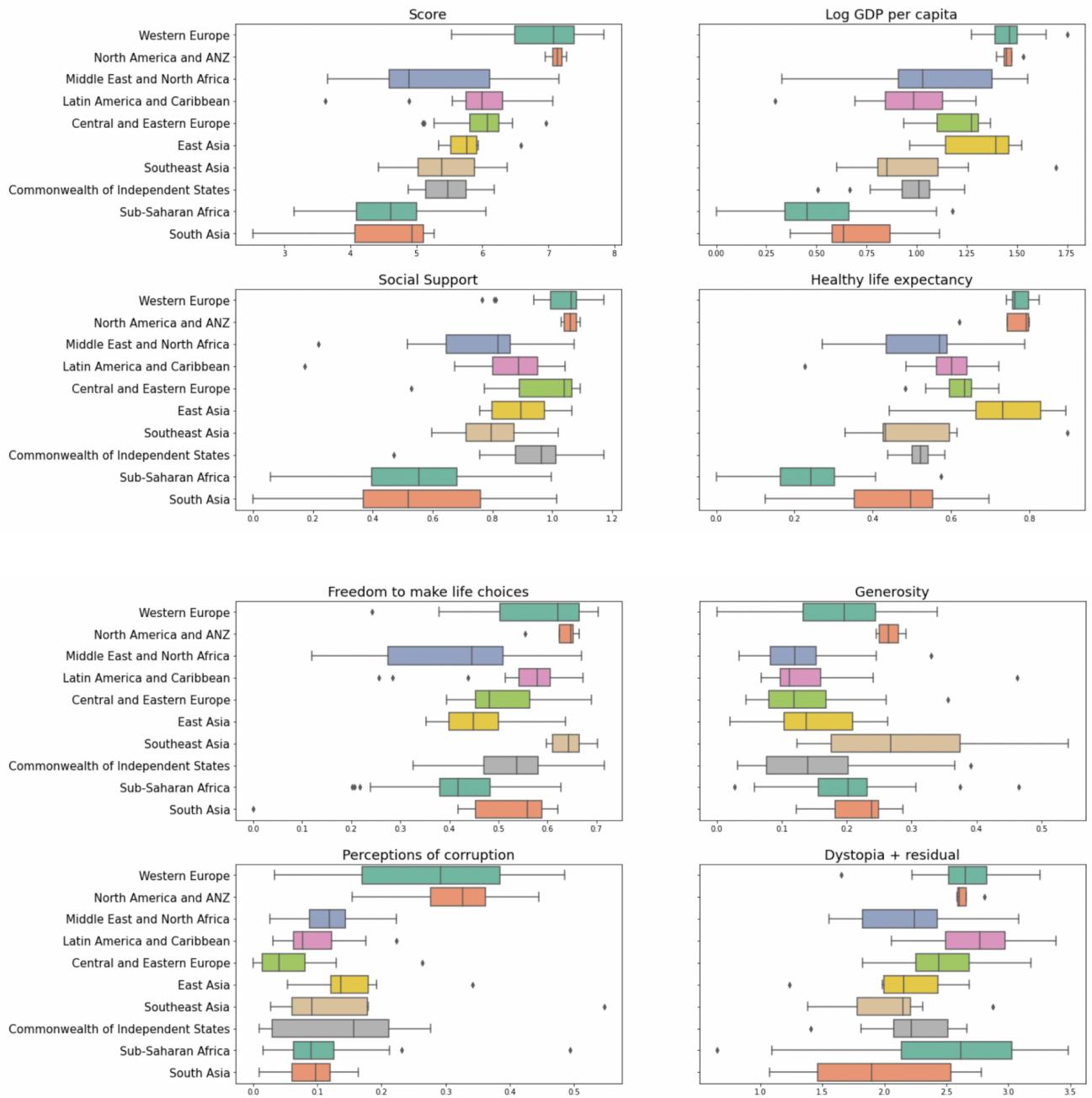
3/ Dispersion des variables explicatives



Observations :

Nous avons utilisé un boxplot pour réaliser une analyse descriptive de données continues afin de visualiser la dispersion des différentes variables qui composent le Ladder score pour les 149 pays du dataset. La dispersion du PIB/habitant est de loin la plus grande, suivie par l'aide sociale et l'espérance de vie. A l'inverse, la générosité et la perception de la corruption semblent être deux variables où cette dispersion est beaucoup moins marquée mais où il existe de nombreuses valeurs extrêmes (outliers).

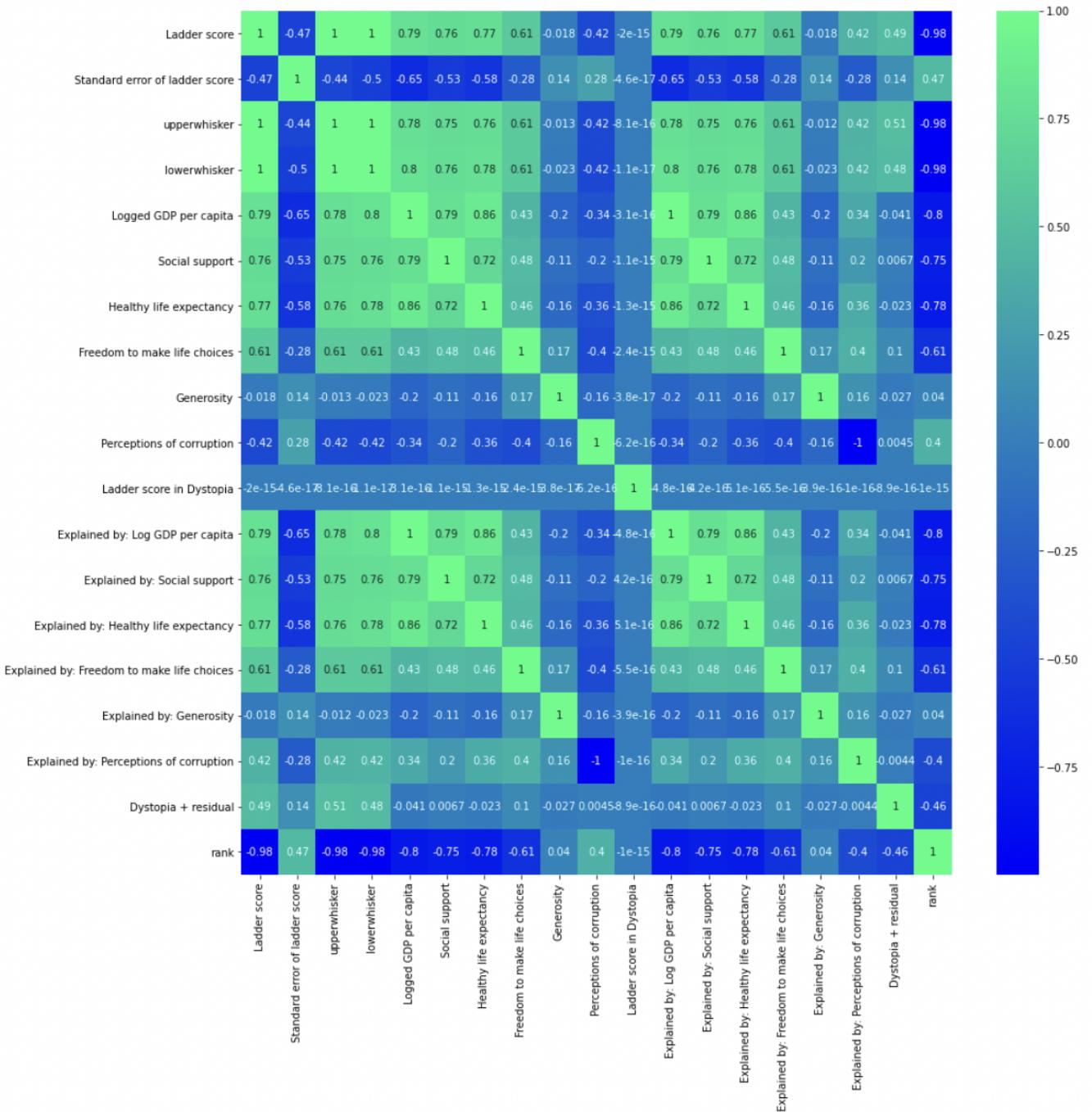
4/ Dispersion des variables par région



Observations :

Cet ensemble de graphiques nous montre l'étendue des variables constitutives du bonheur par région du monde. L'Amérique du nord + Australie + Nouvelle-Zélande représente la région ayant la plus faible étendue sur l'ensemble des graphiques. Ceci s'explique par le faible nombre de pays présents au sein de la région : 4. A l'inverse, la région composée par le plus grand nombre de pays est l'Afrique Subsaharienne avec 36 pays et presque logiquement, son étendue pour chaque variable est très importante.

5/ Corrélation entre variables

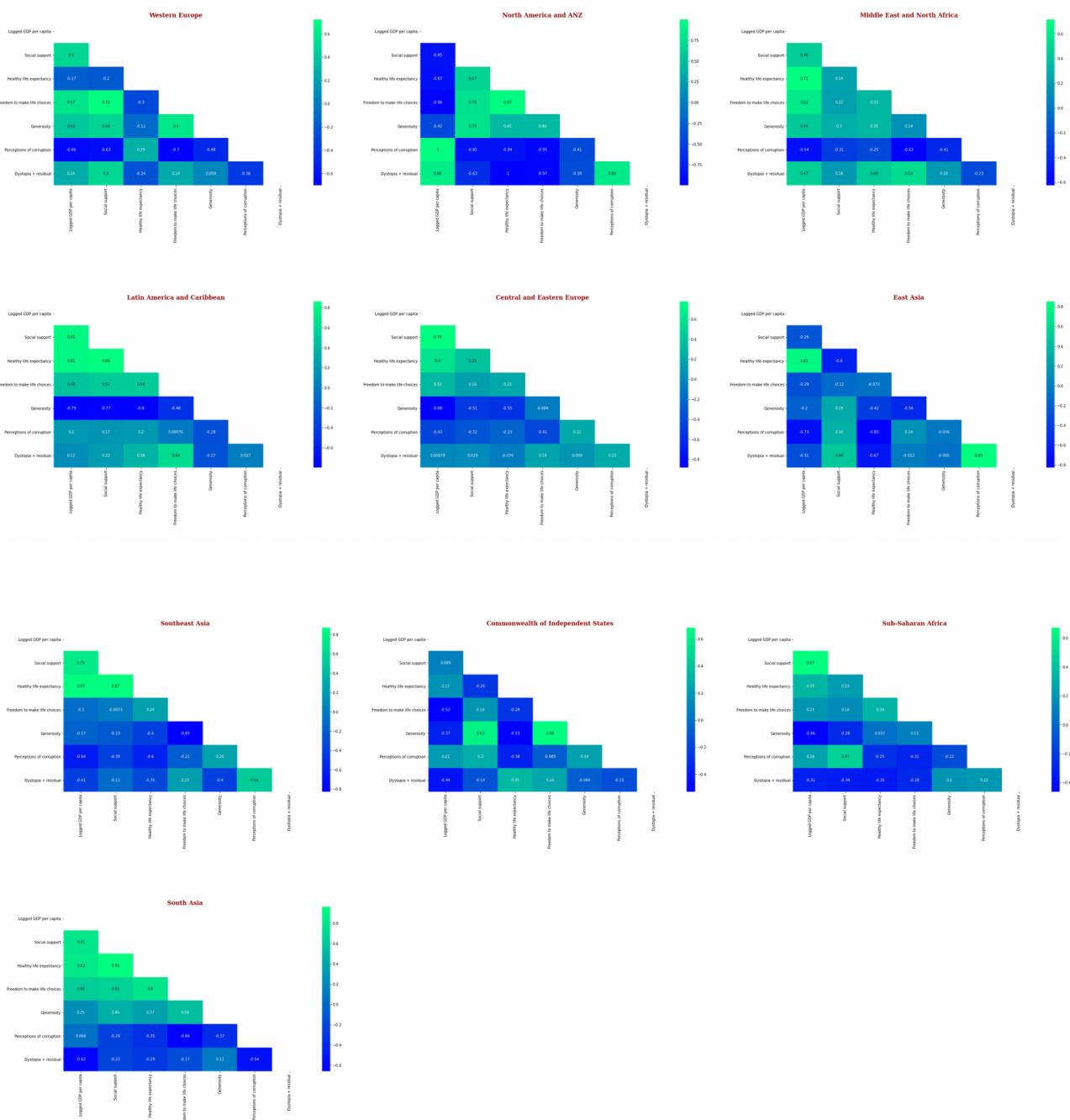


Observations :

Cette heatmap nous indique que les variables, notamment le PIB/habitant, ont la même corrélation avec le Ladder score que les variables transformées comme la variable "Explained by: Log GDP per capita". Nous pouvons donc supprimer l'ensemble des

variables commençant par "Explained by:" de notre Dataset pour pouvoir optimiser les traitements. De même, cette matrice nous confirme que les trois variables les plus corrélées avec le Ladder score sont le PIB/habitant, l'aide sociale et l'espérance de vie. Cependant, ces variables sont fortement corrélées entre elles également, elles ne sont donc probablement pas indépendantes.

6/ Corrélation entre variables par région

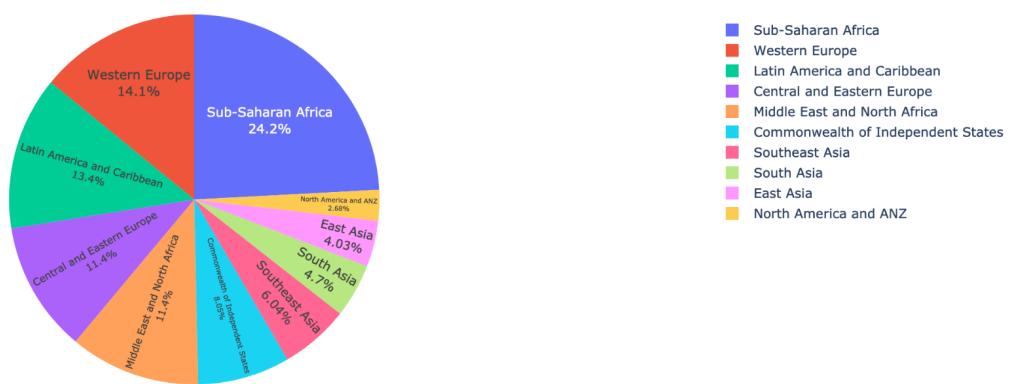


Observations :

Ici, nous avons une heatmap découpée par région du monde. Les variables impactantes ne sont pas toujours les mêmes d'une région à l'autre, mettant ainsi en évidence la subjectivité des critères tout autant que la différence économique et sociale entre chaque région du monde.

7/ Proportion / poids des régions

Répartition des pays par "Regional indicator"

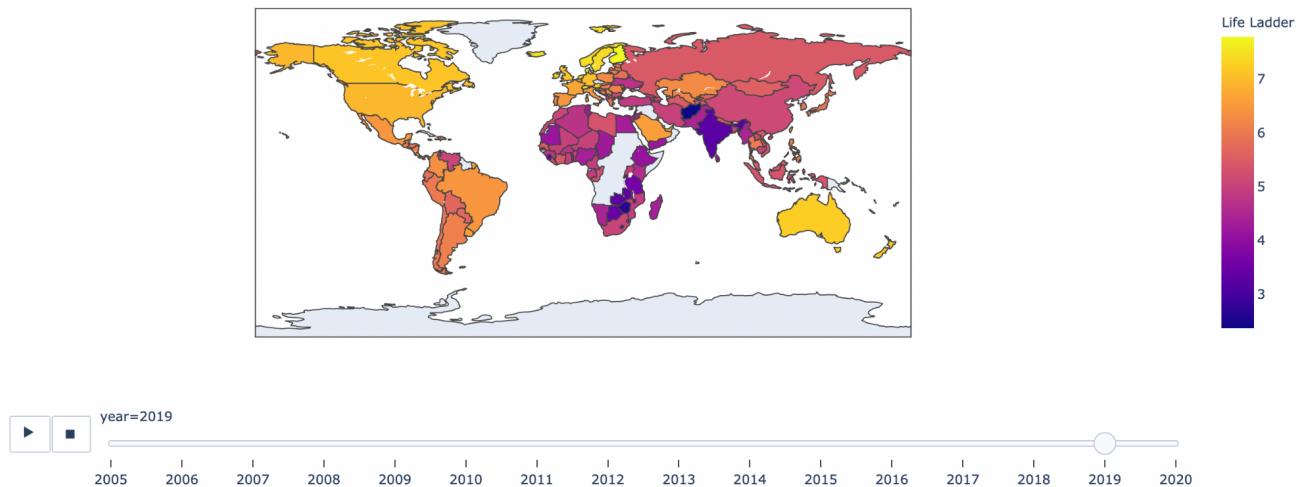


Observations :

Comme indiqué en commentaire du boxplot précédent, l'étendue des variables constitutives du bonheur n'est pas la même d'une région du monde à l'autre. Cela s'explique notamment par l'important écart de nombre de pays au sein de chaque région. Ce graphique en camembert nous permet d'avoir une vue d'ensemble sur le nombre de pays par région du monde et donc leur poids. L'Afrique Subsaharienne, l'Europe de l'ouest et l'Amérique latine + Caraïbes sont les régions les plus importantes qui représentent à elles trois 51.7% des pays. A contrario, l'Amérique du nord + Australie + Nouvelle-Zélande, l'Asie de l'est et l'Asie du sud ne représentent que 11.41% des pays, soit plus de quatre fois moins.

8/ Carte interactive

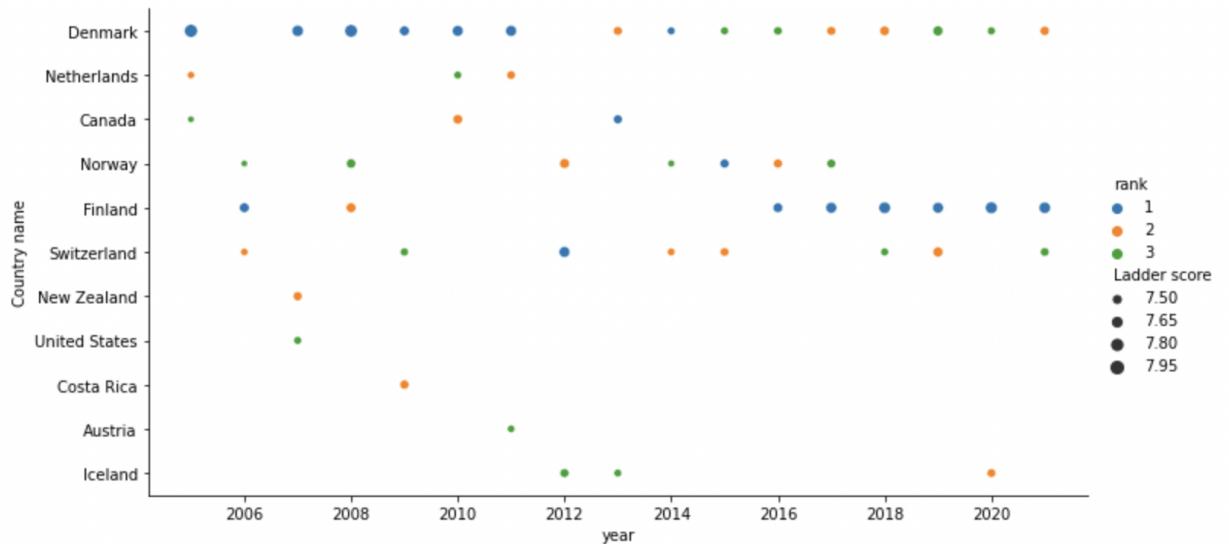
Score "bonheur" dans le monde



Observations :

Cette carte du monde interactive permet de switcher d'années en années, de 2005 à 2020, afin de se rendre compte d'un point de vue plus global, de la répartition des pays les plus heureux, les moins et les intermédiaires ainsi que leur évolution dans le temps. Chaque région du monde semble, à cette échelle, assez bien découpée en fonction du score de bonheur. Se démarquent ainsi l'Europe de l'ouest et l'Amérique du nord + Australie + Nouvelle-Zélande ayant un bon score, suivie par l'Amérique Latine + Caraïbes et Asie de l'est. Se démarquent également l'Asie du sud et l'Afrique Subsaharienne par des couleurs plus sombres, relatives à leur score de bonheur plus bas.

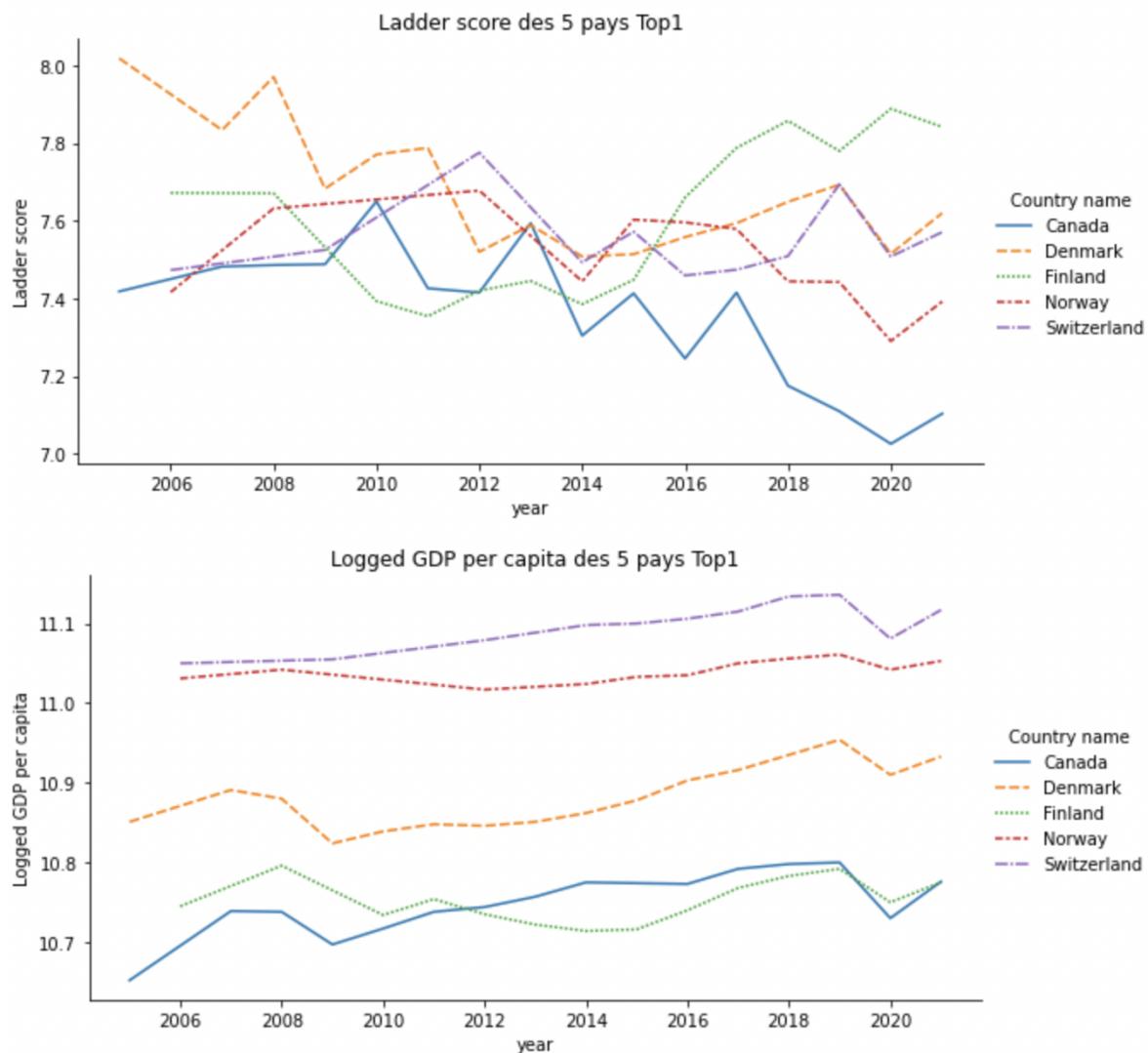
9/ Top 10 de 2005 à 2021

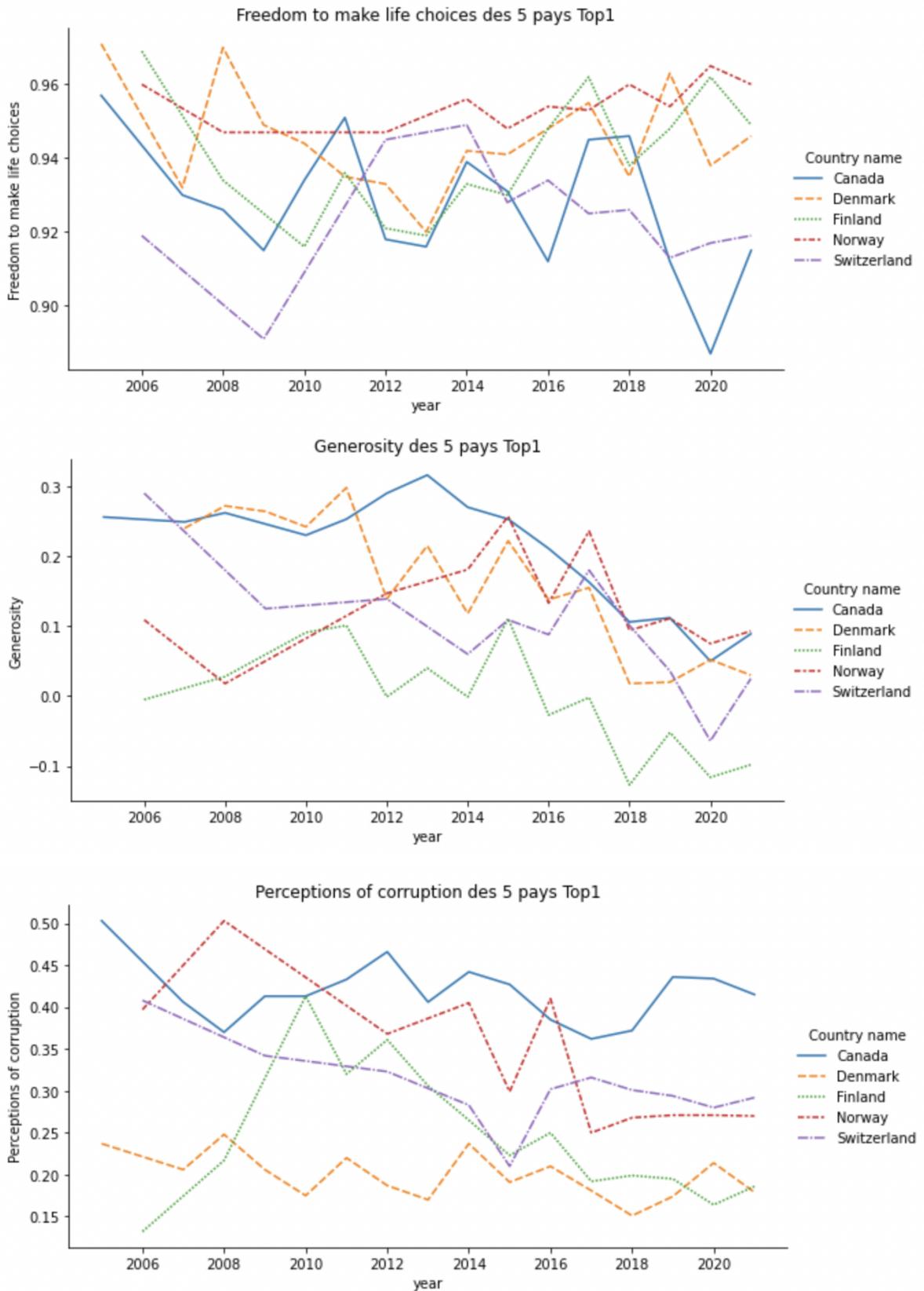


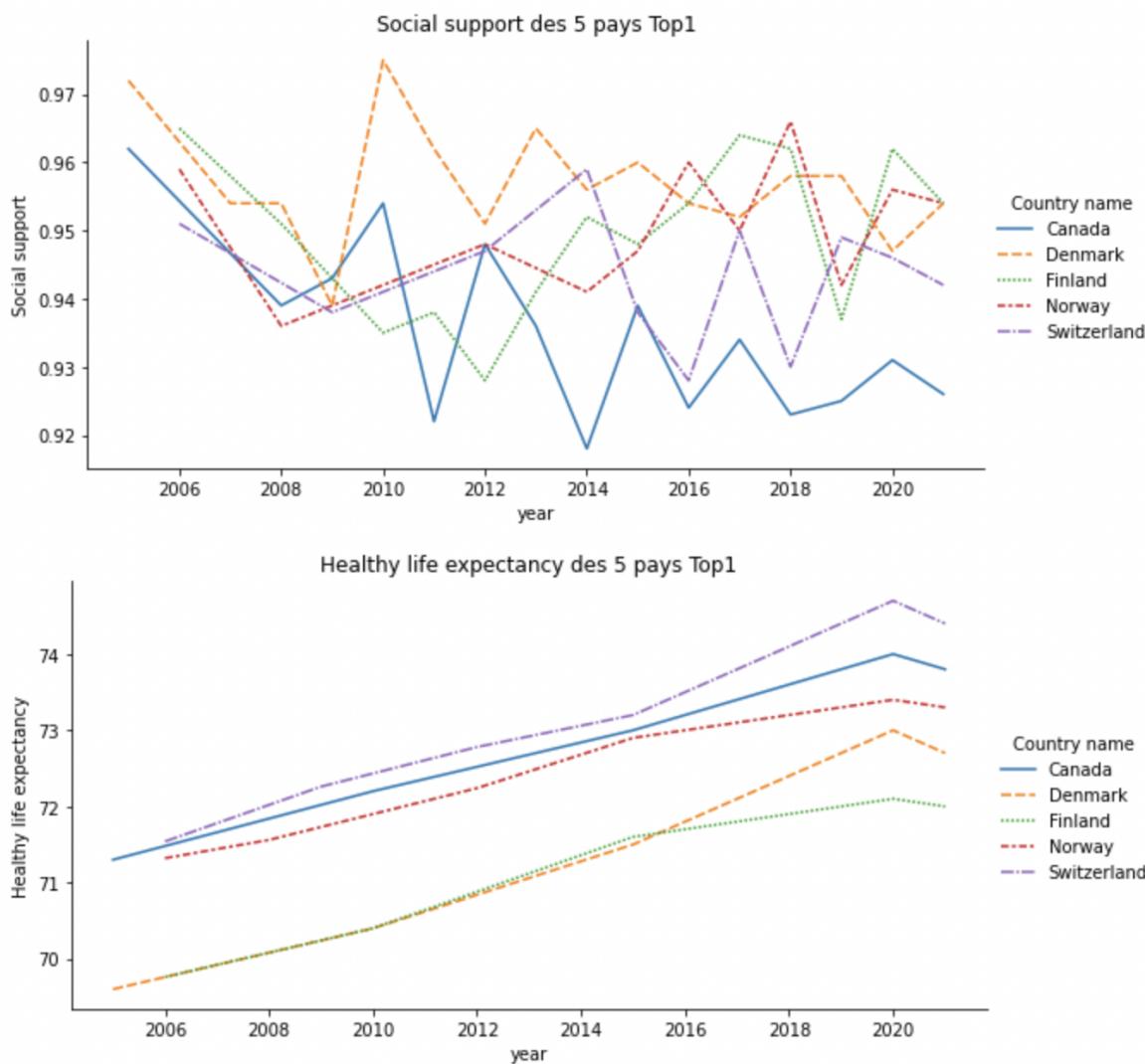
Observations :

Cinq pays se disputent la première place depuis 2005 ; le Danemark, la Finlande, la Suisse, le Canada et la Norvège. La Finlande a su se démarquer dès 2016 pour arriver en tête du classement. Également, depuis 2014, le top 4 est uniquement occupé par des pays d'Europe de l'ouest où la Finlande et le Danemark semblent être les pays les plus heureux.

10/ Evolution du ladder score et des variables explicatives du top 5







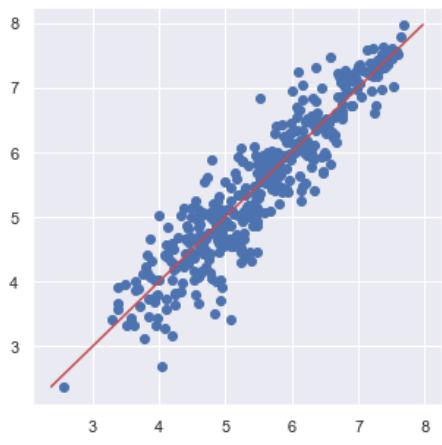
Observations :

Depuis six ans, la Finlande est en tête de ce classement alors que la Suisse présente de meilleurs résultats sur 2 des 3 variables les plus corrélées au Ladder score (PIB/habitant et espérance de vie). On peut facilement observer que pour chacun des cinq pays, les variables PIB/habitant et espérance de vie sont assez linéaires et augmentent d'années en années à l'exception de l'année 2020, marquée par la crise sanitaire qui perdure encore aujourd'hui. A l'inverse, les autres variables alternent entre rebond et rechute de façon cyclique. La générosité quant à elle, agit de même mais elle a tendance à diminuer d'années en années mais connaît un rebond récent.

Modélisation

Nous avons opté pour deux approches différentes, une régression pour prédire le Ladder score, et une classification afin de découper nos pays en deux catégories : pays heureux et non-heureux. Nous avons opté comme métrique de performance principale l'accuracy pour comparer nos modèles. Nous avons donc utilisé et comparé plusieurs modèles, notamment un Support Vector Machine (SVM), une régression linéaire simple et une régression linéaire régularisée avec Lasso pour prédire le Ladder score. Pour la catégorisation/classification, nous avons opté pour un Support Vector Machine (SVM), une méthode des k plus proches voisins (KNN) et des forêts aléatoires (Random Forest). Nous avons également optimisé ces modèles avec les meilleurs paramètres et hyperparamètres afin d'obtenir la meilleure performance possible.

Les 6 variables sont nécessaires au bon fonctionnement des modèles (essais effectués via SelectFromModel).

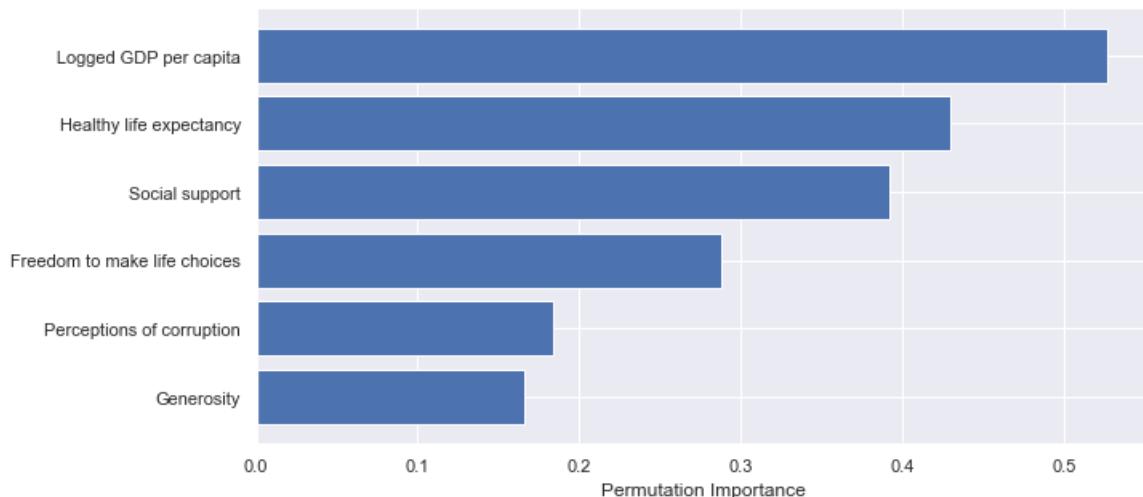


Pour la prédiction du Ladder score, nous avons retenu le modèle SVM, étant le modèle présentant le meilleur score sur l'échantillon test.

La recherche des meilleurs paramètres via la GridSearchCV a permis de passer d'un score de 0,81 à 0,85 sur l'échantillon test (`{'C': 5, 'gamma': 0.5, 'kernel': 'rbf'}`).

Coef de détermination du modèle : 0.92
Coef de détermination obtenu par cv : 0.83
Score test : 0.85

L'analyse de la Permutation Importance de notre modèle SVM nous confirme que les variables les plus importantes dans ce modèle sont le PIB par habitant, l'espérance de vie et l'aide sociale.



Pour la catégorisation, nous avons opté pour un Random Forest. C'est le modèle qui nous a donné la meilleure précision mais c'est également celui possédant l'exécution la plus rapide. La recherche des meilleurs paramètres pour le SVM prenait plusieurs heures pour un résultat équivalent en termes d'accuracy au Random Forest. Concernant le KNN testé, celui-ci a obtenu un score légèrement moins élevé de 2 points, soit 86% contre 88% pour le Random Forest.

A titre informatif, voilà les résultats que nous avions obtenu avec la RandomForest:

		Classe prédictive		1	2
		Classe réelle			
		1	184	14	
		2	35	174	
		precision	recall	f1-score	support
1		0.84	0.93	0.88	198
2		0.93	0.83	0.88	209
accuracy				0.88	407
macro avg		0.88	0.88	0.88	407
weighted avg		0.88	0.88	0.88	407

Nous gardons, comme modèle final, le SVM présenté précédemment car la classification nous paraissait assez peu pertinente (la séparation des résultats entre deux catégories "heureux" et "non heureux" n'étant pas vraiment représentative).

Conclusion

Ce projet nous a permis de confronter nos connaissances et compétences récemment acquises à un cas pratique.

Nous sommes passés par les étapes suivies durant cette formation; l'exploration de données, le data cleaning, la datavisualisation avec différentes libraires (matplotlib, seaborn, plotly), le choix et la comparaison des modèles de machine learning.

Durant ce projet, chacun a pu apporter son expertise liée à son domaine d'expérience passé. Cela nous a permis de travailler conjointement tout en s'aidant pour le mener à bien.

L'étude de ce sujet nous a permis de prendre conscience du biais de cette étude où le bonheur d'un pays s'appuie très largement sur sa richesse. Ainsi, les pays moins riches où le bonheur peut être plus élevé au sens strict du terme, seront pénalisés dans cette étude où la pondération des variables explicatives leur porte préjudice. Le graphique "Permutation Importance" présenté précédemment met en exergue l'importance de 3 variables : le PIB/habitant, l'aide sociale et l'espérance de vie. Ce sont-là trois variables qui sont généralement plus élevées dans les pays riches et le score de bonheur est donc tronqué pour les pays pauvres.

Nous aurions souhaité, à l'aide d'outils de web scraping et text mining, analyser les différents médias de chaque pays afin de comparer le bonheur des habitants d'un pays versus la situation décrite dans ces médias.

Nous avons en ce sens développé une fonction récupérant les informations de chaque pays sur Wikipédia, notamment le contexte politique et les données socio-économiques. Cela aurait permis avec une recherche de l'actualité de chaque pays, une analyse complémentaire.

Annexes

Diagramme de GANTT :

Taches	AOUT										SEPTEMBRE																	OCTOBRE																				
	23	24	25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	1	2	3	4	5	6	7		
Exploration des données																																																
Choix des objectifs du projet																																																
Data cleaning																																																
Dataviz																																																
1ère itération																																																
2nde itération																																																
3ème itération																																																
Rédaction rapport																																																
Création streamlit																																																
Verification finale (codes + rapport)																																																
Entraînement à la soutenance																																																
Soutenance																																																

Bibliographie :

- ❖ <https://worldhappiness.report/ed/2021/>