

HapPy Factory

Explication du score de bonheur national déclaré par ses composantes supposées

Peut-on expliquer les résultats du sondage annuel et international World Happiness Report ? Avec les réponses aux questions complémentaires ou avec des données objectives annexes, nous exploitons différentes méthodes d'analyse de données et de régression linéaire pour sélectionner les variables explicatives et établir des interprétations sur le bonheur.

Oussama BAKKOURY
Arthur MAROT
Hugo ROBLEDO-GRUEL

| | |
|---|-----------|
| Introduction | 2 |
| Analyse des données | 3 |
| Analyse des données initiales | 3 |
| Analyse des données manquantes | 4 |
| Analyse des corrélations | 5 |
| Analyse des tendances dans les régions du monde | 7 |
| Analyse des nouvelles variables exploitées | 11 |
| Démarche et sélection des variables | 14 |
| Régression Linéaire multiple | 14 |
| Représentation sur dimensions réduites | 15 |
| Modèle de classification | 16 |
| Ajout de variables complémentaires | 18 |
| Modélisation avec les variables ajoutées | 18 |
| Modèle retenu et interprétation | 21 |
| Modèle de Lasso | 21 |
| Modèle à inertie | 23 |
| Capacité de prédiction | 24 |
| Conclusion | 26 |
| Sources | 27 |
| Annexes | 28 |

Introduction

Le World Happiness Report est une publication d'une antenne de l'ONU reposant sur les résultats du Gallup World Poll, un sondage annuel et international auprès d'un millier d'habitants par pays. L'objectif de ce sondage est de mesurer un score de bonheur moyen local sur la base des déclarations des sondés, qui sont également interrogés sur leur sentiment de liberté et de soutien social, leur perception de la générosité et de la corruption et des sentiments positifs et négatifs récents. Toutes ces questions ont été posées de sorte à les rapporter sur une échelle de 10.

Ce rapport prétend faire l'état des lieux du "bonheur" dans le monde avec chaque année des thèmes abordés pour mesurer et favoriser le progrès mondial. Avec 100 à 140 pays interrogés par an depuis 2007 et jusqu'à 2021, une grande partie des états sont représentés (l'ONU en reconnaît 197 aujourd'hui), bien que certains n'aient pas été interrogés tous les ans, soit parce que ces pays sont anciens ou nouveaux, parce qu'ils sont dans une situation instable ou encore parce qu'ils pèsent moins à l'échelle du monde.

A ces données, ont été ajoutés un indicateur du PIB par habitant dans chaque pays sondés (c'est le logarithme qui est retenu car il permet d'ignorer la croissance exponentielle du PIB), l'espérance de vie moyenne des habitants, et la région du monde auquel appartient chaque pays. Des données qui ont la particularité de pas être subjectives, contrairement à celles du sondage.

Nous nous sommes fixé pour objectif d'expliquer le score de bonheur moyen d'un pays avec une modélisation reposant sur d'autres variables observées. Sans la prétention de pouvoir prédire ce score, nous visons à décomposer le score de bonheur par des composantes ayant un poids plus ou moins important, positif ou négatif, et à en tirer des conclusions.

Dans cette optique, nous avons fait le choix d'ajouter des données objectives supplémentaires pour la modélisation. Nos intuitions nous ont amenés à introduire des notions d'inégalités, de religion, de régime politique, de corruption mesurée et de démographie; des données que nous avons effectivement retenues à des fins de modélisation.

Grâce à l'analyse de l'ensemble de ces données nous avons progressivement sélectionné des composantes permettant de modéliser le score de bonheur local. La modélisation a été réalisée via plusieurs méthodes mathématiques jusqu'à en retenir la plus pertinente : une régression linéaire dont nous analysons les rouages pour en tirer des interprétations sur les composantes du bonheur.

Cette démarche nous a amené à obtenir un modèle prédictif relativement robuste qui met en évidence des valeurs humaines universelles, et surtout qui contredit une première vision simpliste mettant en évidence le poids majeur du PIB dans le score de bonheur. Un modèle qui permet d'offrir une certaine vision de la recette du bonheur. C'est l'origine du nom de ce projet : HapPy Factory.

Analyse des données

Analyse des données initiales

La table initiale de notre jeu de données a 12 colonnes et 1949 lignes concernant un pays une année. Ce sont des données numériques à l'exception de la région du monde à laquelle appartient le pays. Nous avons 166 pays au total, pour 10 régions et 16 années représentées, avec des valeurs manquantes assez rares. Les variables avec le plus de valeurs manquantes sont "Perceptions of corruption", "Generosity" et "Healthy life expectancy at birth".

Number of duplicated data: 0

| | Column | Dtype | Number of values | NA values | Unique values |
|----|----------------|---------|------------------|-----------|---------------|
| 0 | Country | object | 1949 | 0 | 166 |
| 1 | Year | int64 | 1949 | 0 | 16 |
| 2 | Life Ladder | float64 | 1949 | 0 | 1553 |
| 3 | LogGDP | float64 | 1913 | 36 | 1500 |
| 4 | SocialSupport | float64 | 1936 | 13 | 455 |
| 5 | LifeExpectancy | float64 | 1894 | 55 | 828 |
| 6 | Freedom | float64 | 1917 | 32 | 535 |
| 7 | Generosity | float64 | 1860 | 89 | 609 |
| 8 | Corruption | float64 | 1839 | 110 | 572 |
| 9 | PosAffect | float64 | 1927 | 22 | 431 |
| 10 | NegAffect | float64 | 1933 | 16 | 374 |
| 11 | Region | object | 1949 | 0 | 10 |

Image 1 : Tableau d'information sur les colonnes

On constate 4 échelles de valeurs :

- dans les 2000 pour les années
- entre 50 et 100 pour l'espérance de vie "Life Expectancy"
- 5 à 12 pour Life Ladder et LogGDP
- en pourcentage pour les autres questions du sondage

Toutes les données ont une répartition relativement équilibrée (pas de valeurs aberrantes).

A surveiller la notion de générosité (Generosity : min -0.33 à q3 0.09 puis max loin à 0.69) et la perception de la corruption (Corruption : min loin 0.03 puis q1 0.7 à max 0.98); Celles-ci auraient pu justifier une normalisation, bien que cela n'ait pas été nécessaire.

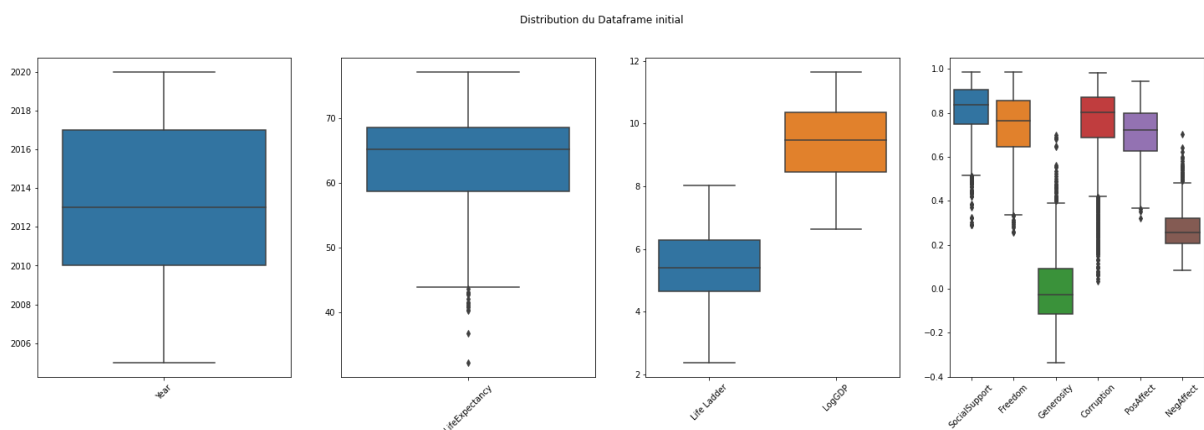


Image 2 : Boîte à moustache des valeurs numériques du dataframe

Analyse des données manquantes

Au global, les données manquantes sont rares mais certaines années le nombre de pays représentés était beaucoup plus faible, l'année 2005 notamment ne concerne qu'une poignée de pays.

Il convient donc d'analyser l'impact des années et des pays représentés pour prévenir un éventuel biais. On peut envisager de ne considérer que les années avec au moins 100 pays (2007 à 2019); à savoir que l'ONU reconnaît 195 pays.

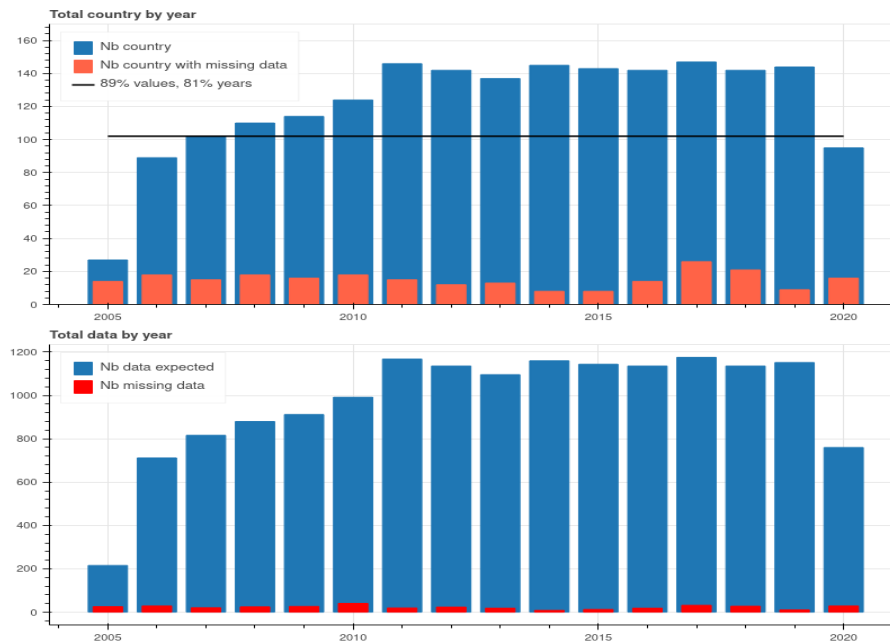


Image 3 : Graphiques nombre de pays et nombre de données par an avec une indication sur les valeurs manquantes

Les régions qui ont le plus de données manquantes sont "East asia" et "middle East countries" :

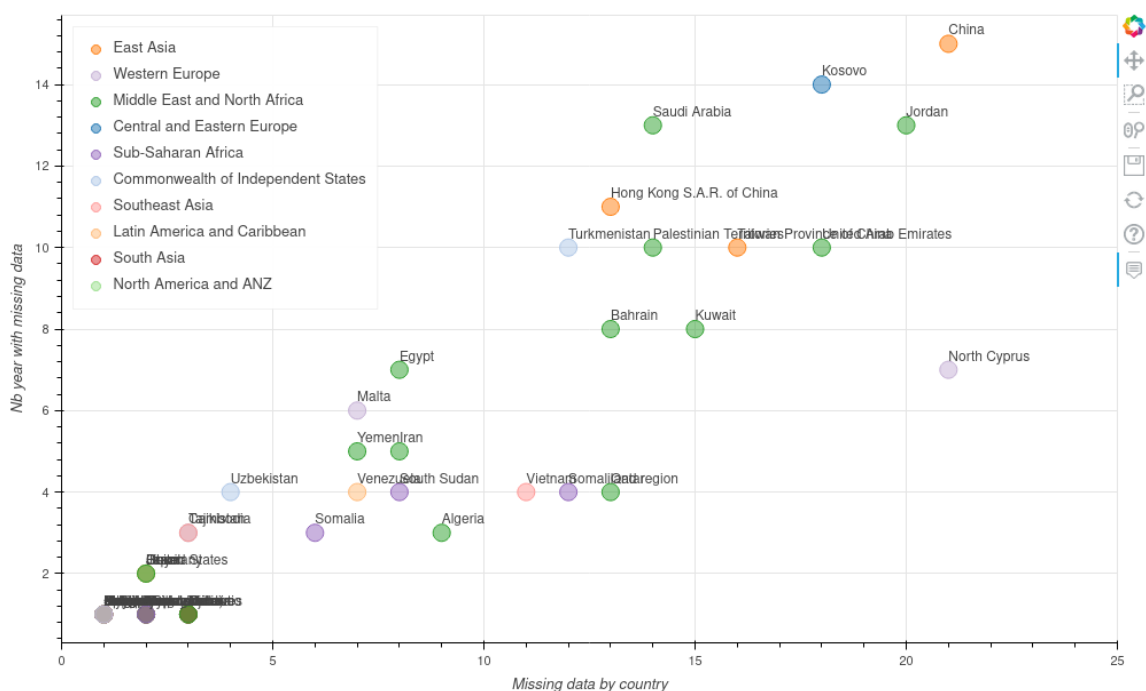


Image 4 : Nuage de point du nombre de valeurs manquantes par rapport au nombre d'année avec des valeurs manquantes par pays

Analyse des corrélations

Les Variables les plus corrélées au score de bonheur "Life Ladder" sont,
de façon positive :

LogGDP : 0.79, LifeExpectancy : 0.74, SocialSupport : 0.71, Freedom : 0.53 et PosAffect : 0.53;

et de façon négative :

Corruption : -0.43 et NegAffect : -0.3.

La générosité est peu corrélée.

Des variables ont aussi des corrélations fortes entre elles au delà du score de bonheur :

LogGDP et SocialSupport : 0.69, LifeExpectancy et SocialSupport : 0.62, Freedom et PosAffect : 0.61,
voir très forte entre elles, LifeExpectancy et LogGDP : 0.85

Les informations que nous apportent ces données sont donc peut-être dédoublées, au sens où elles décrivent les mêmes phénomènes locaux, partiellement ou intégralement, il faudra en tenir compte dans le choix de nos modèles.

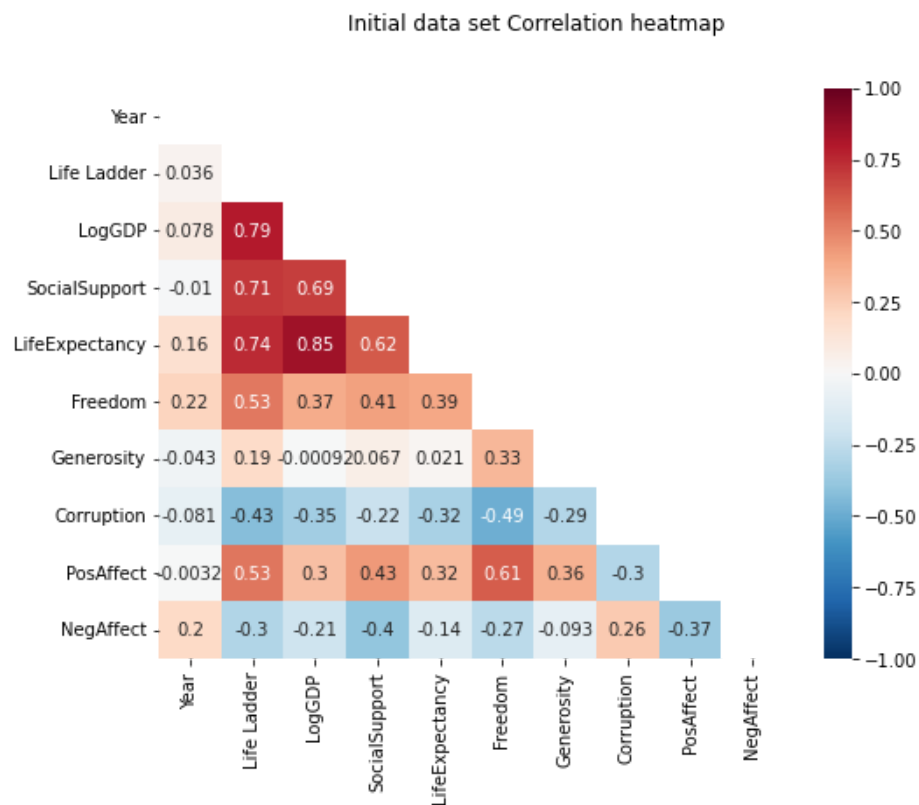


Image 5 : Matrice de corrélation de la data frame initiale

Suite à ces observations, nous avons fait le choix d'éliminer simplement les données manquantes. Le jeu de données "whr_NoNA" à 1708 lignes et 155 pays, le nombre d'années reste inchangé. Nous conservons 88% des lignes (-241) et 93.4% des pays (-11) du jeu de données initial. La dispersion des données reste très proche avec et sans les valeurs manquantes :

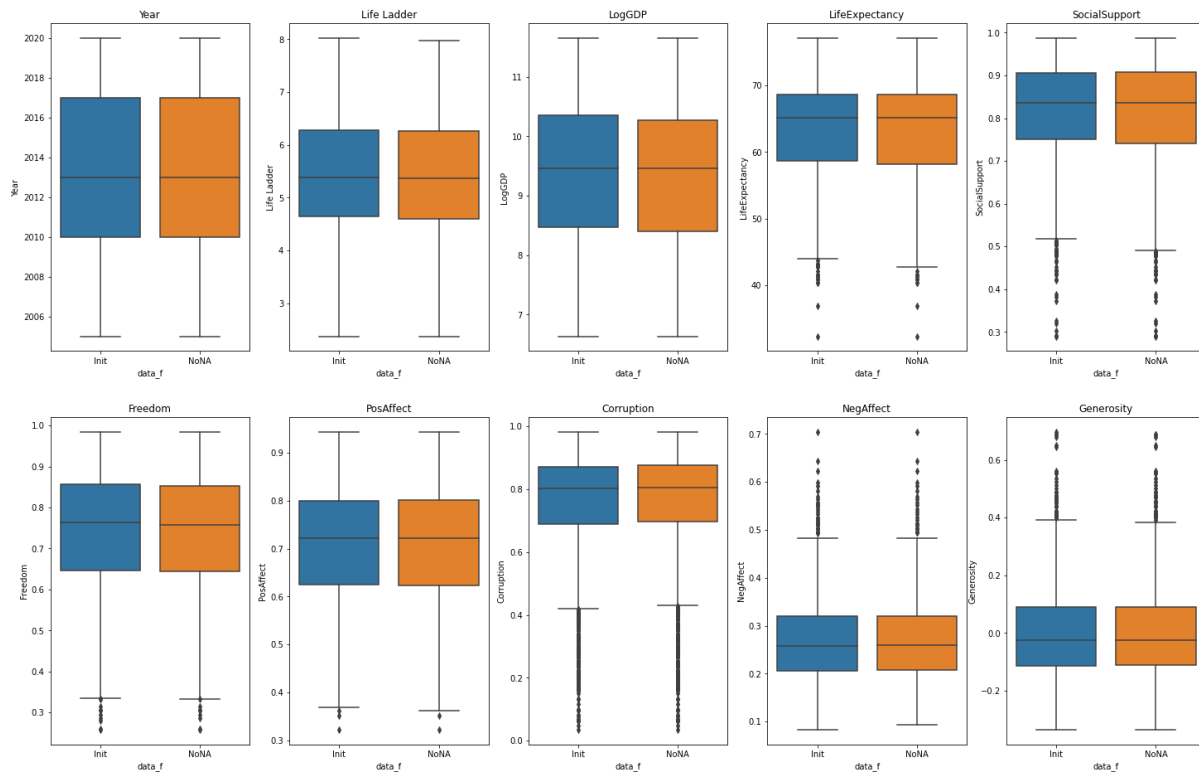


Image 6 : Boîte à moustache de toutes les variables numérique pour comparer les données initiales et les données sans NA

Les scores de corrélations restent aussi très similaires :

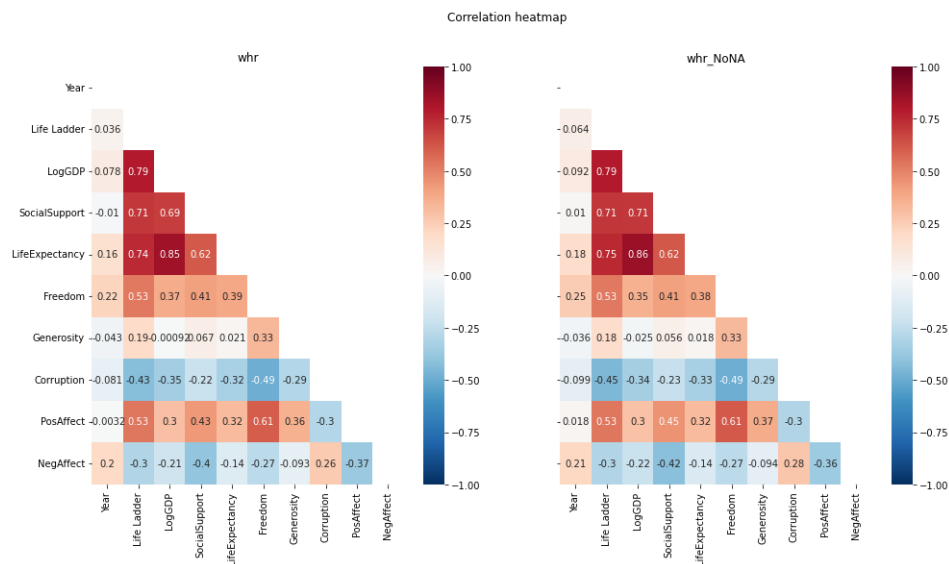
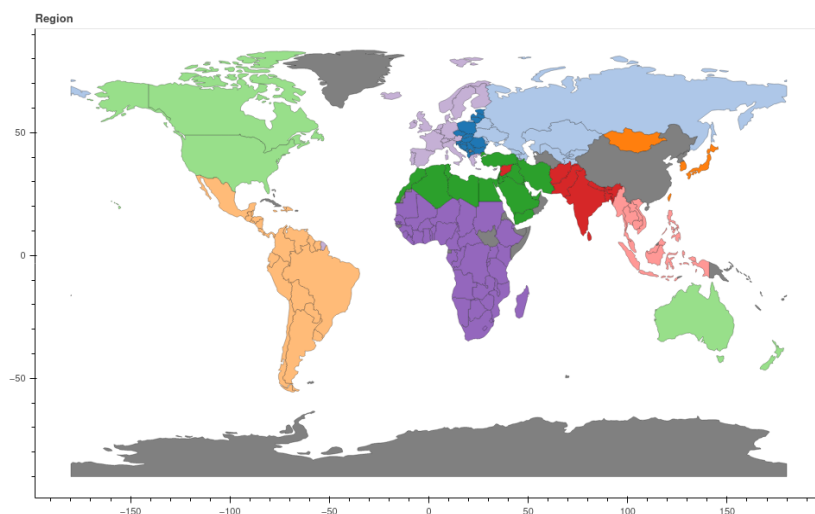


Image 7 : Matrice de corrélation de toutes les variables numérique pour comparer les données initiales et les données sans NA

C'est pourquoi nous avons retenu ces données pour les étapes suivantes de notre étude. Suivant les résultats, nous envisagions de revenir sur cette décision, soit réintégrer des lignes en retirant certaines variables, soit estimer des valeurs manquantes, mais cela n'a pas été nécessaire.

Analyse des tendances dans les régions du monde

Voici le découpage des 10 régions du monde associées à chaque pays :



Amérique du Nord et Australie en **vert clair**, Amérique latine en **orange clair**, Europe de l'Ouest en **violet clair**, Europe centrale et de l'Est en **bleu foncé**, Moyen-Orient et Afrique du Nord en **vert foncé**, Afrique sub-saharienne en **violet**, Union de la Russie et Asie central en **bleu clair**, Asie du Sud en **rouge**, Asie du Sud-Est en **rose**, Asie de l'Est en **orange**.

Image 8 : Carte avec un code couleur par région

Nous avons observé le score de bonheur par année et par pays sur une carte avec un code couleur :

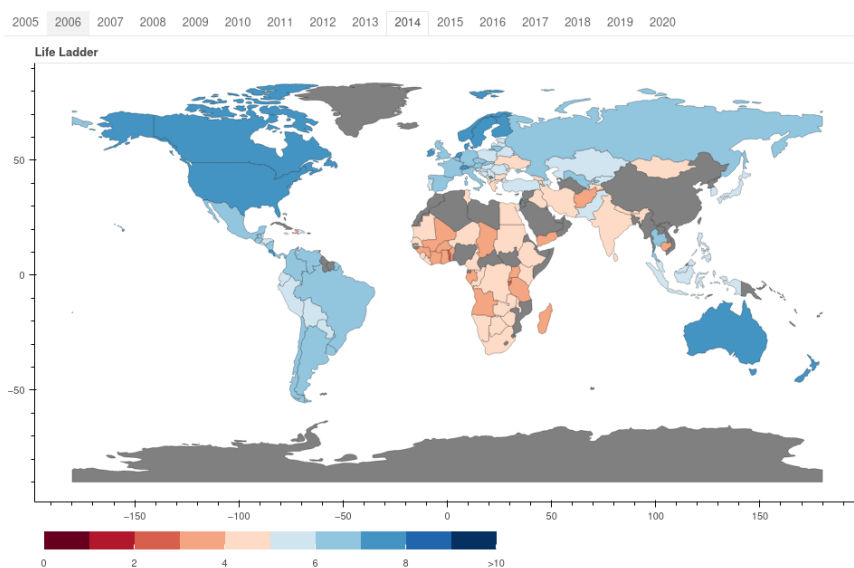


Image 9 : Carte avec le score du Life Ladder en échelle de couleur de l'année 2014

Les pays Scandinaves, de l'Amérique du nord, l'Australie et la Nouvelle Zélande ont les scores les plus élevés et ils semblent stables. Ensuite, nous avons les pays d'Europe de l'Ouest et les pays d'Amérique latine qui semblent majoritairement bons et stables (avec quelques exceptions). Les pays d'Europe centrale et de l'Est, Moyen-orient et Afrique du Nord, union de la Russie et Asie central, Asie du Sud, Asie de l'Est semblent avoir des scores moyens et qui évoluent un peu au fil des années. Enfin, les deux régions avec des pays dont les scores sont les plus bas viennent de l'Asie du Sud, et de l'Afrique sub-saharienne.

On peut faire des observations similaires avec un graphique de la médiane du score de bonheur en fonction de l'écart-type pour chaque pays.

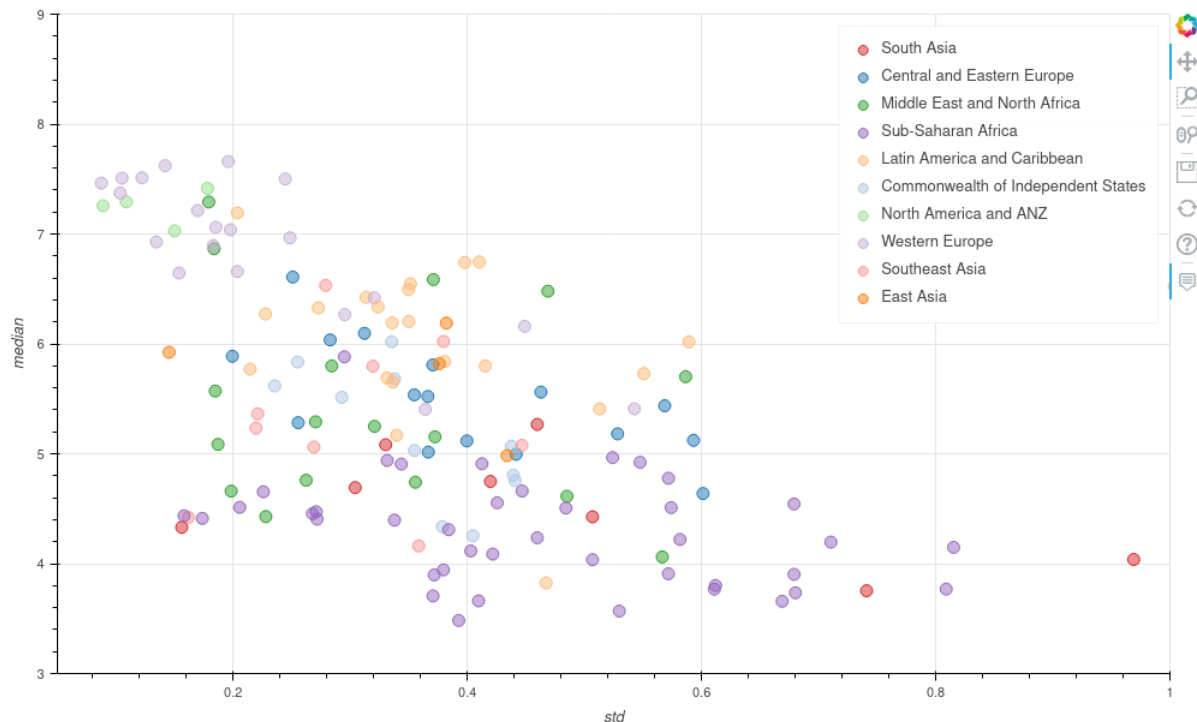


Image 10 : Médiane du score de bonheur en fonction de l'écart-type pour chaque pays

Nos constats précédents sont confirmés mais on peut également s'intéresser aux pays avec plus ou moins d'évolution (écart-type) et/ou avec un score faible ou élevé. On voit que le score médian de bonheur dépend bien des régions, mais c'est moins évident pour l'écart-type. On peut donc conclure que plus le score est élevé, plus il y a de chances que la variation soit faible au fil du temps.

Exemples :

Le plus stable et le plus haut : Netherlands

Le plus stable et le plus bas : Sri Lanka

Le moins stable et le plus haut : Non existant

Le moins stable et le plus bas : Syrie

Distribution des variables par région grâce à des boîtes à moustaches :

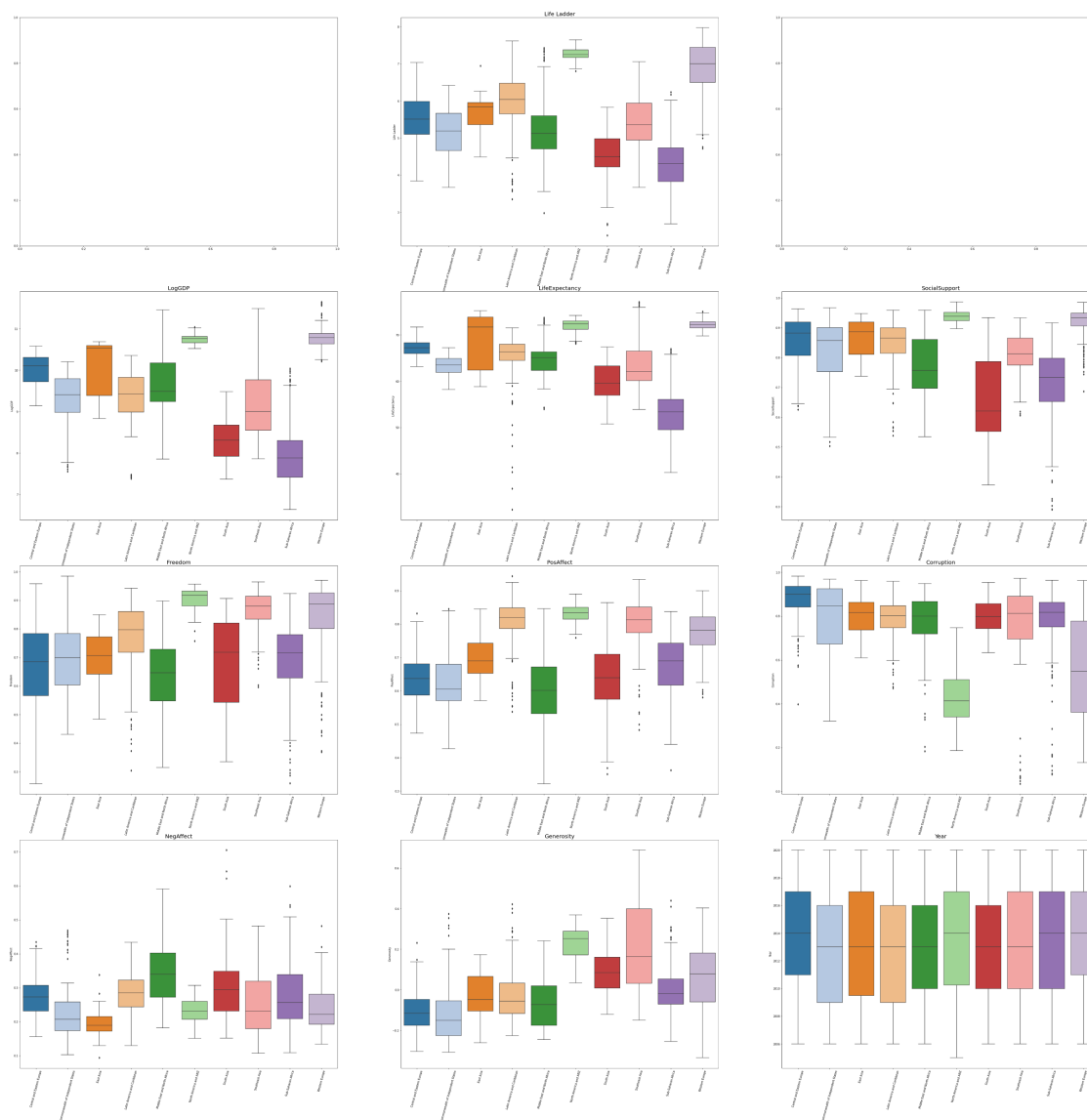


Image 11 : Boîte à moustache de toutes les variables numériques pour comparer les régions

Il y a de grandes similarités entre la distribution des variables score de bonheur et PIB par région :

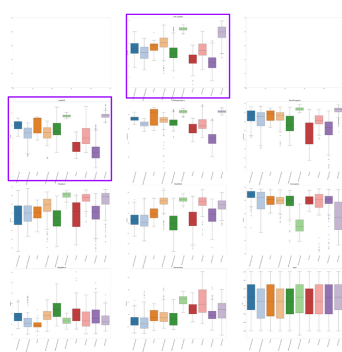


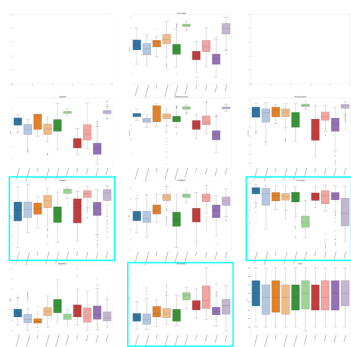
Image 12 : Aide à la lecture de l'Image 11

Les pays d'Europe ont des PIB très similaires mais leurs scores sont très variés. A l'exact opposé, dans les deux régions d'Asie, avec des PIB très variés on observe des scores plus proches. Les variables sont très dispersées au sein de certaines régions, cependant il convient de noter que ces régions ont des échelles très différentes, en particulier le Moyen-Orient et Afrique du Nord. Enfin, la région Amérique latine comporte une série de pays avec un score particulièrement faible.

De manière générale, la distribution annuelle par région est relativement similaire. Toutes les années peuvent donc être étudiées indifféremment, et nous les avons utilisées pour augmenter le volume de données d'entraînement plutôt qu'en temps que variable utile.

L'analyse des distributions des variables confirme l'hétérogénéité entre les régions, il y a une opposition Nord/Sud (hors Amérique latine et Océanie) qui n'est pas surprenante.

En Europe, on remarque une forte hétérogénéité des valeurs dans chaque variable, il serait donc intéressant de différencier ces pays selon d'autres critères, nous pensons au régime politique en particulier. Tandis qu'en Europe de l'ouest, on constate un pessimisme marqué, ceci confirme l'importance d'un paramètre comme la culture comparé à d'autres continents. Au sein de l'Europe on pourrait essayer de comparer les pays du nord et ceux du sud (dits "latins").



Plus surprenant, la perception de la liberté n'est pas réservée aux pays occidentaux. Et les sentiments positifs sont plus forts dans les pays d'Amérique latine.

Dans l'Asie du Sud, la perception de liberté est plutôt bonne, le sentiment de corruption est moyen et il y a une forte générosité entre individus. Pourtant le bonheur ne suit pas.

A l'inverse, l'Amérique latine a une corruption forte et il y a peu de générosité. Pourtant les populations semblent plutôt heureuses.

Image 13 : Aide à la lecture de l'Image 11

Certaines corrélations semblent évidentes, mais l'étude des régions met en évidence une limite des variables analysées à ce stade. Des variables semblent manquées afin de mieux comprendre le score de bonheur observé au sein des pays et des régions. Le PIB, notamment, semble regrouper trop d'informations. Nous avons donc ajouté de nouvelles variables pour mieux comprendre ces phénomènes.

Analyse des nouvelles variables exploitées

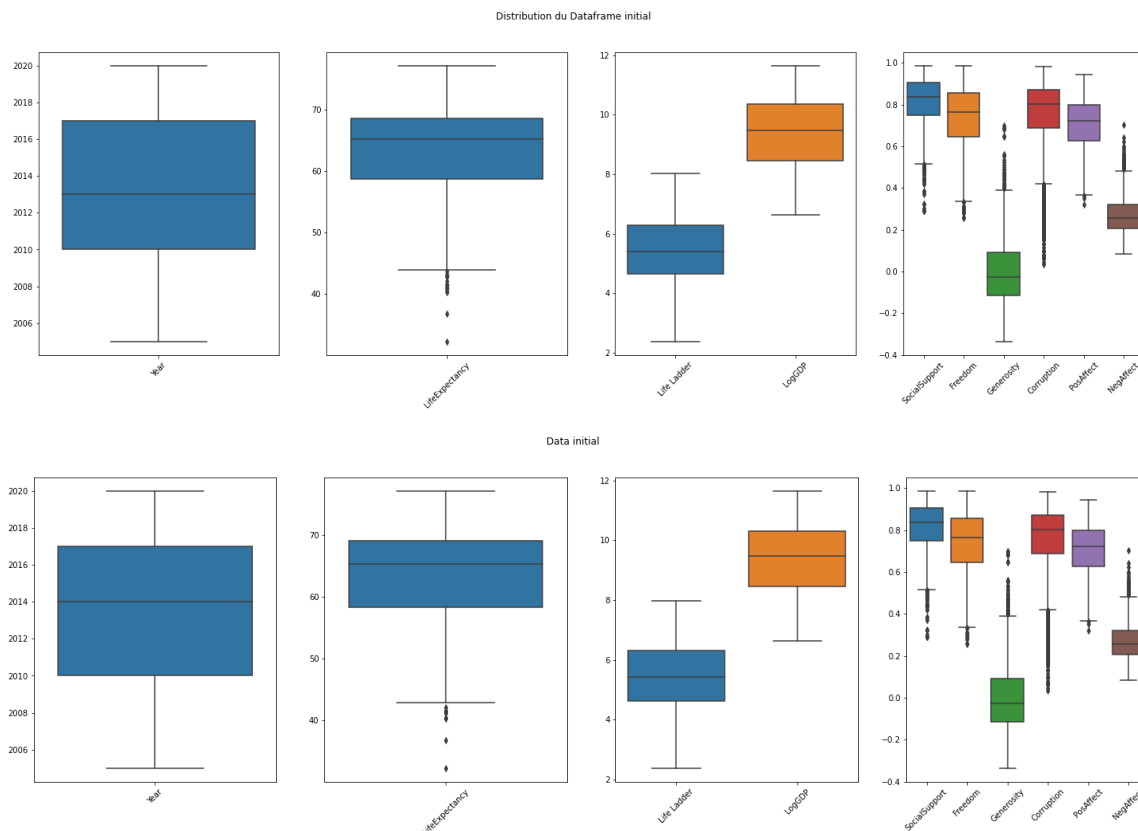
Notre volonté est de décanter l'impact du PIB avec des composantes culturelles. Grâce aux analyses précédentes nous avons retenu une série de variables complémentaires pour expliquer le score de bonheur national :

- la population du pays : nombre d'habitant
- la superficie du pays : surface en km2
- la densité de population : nombre d'habitant / la superficie du pays
- le taux de mortalité infantile
- l'indice d'inégalités de l'année 2021 (coefficient de Gini)
- la corruption mesurée : une donnée plus objective que celle du sondage ressenti
- la croissance annuelle de la population
- la classe de régime politique (démocratie, démocratie imparfaite, hybride, dictature)
- la part de la population croyante pour chaque religion: Christianisme, Islam, Sans-religion, Hindouisme, Bouddhisme, Religions traditionnelles, Autres, Judaïsme

Une fois les variables ajoutées nous refaisons une analyse de notre nouveau jeu de données.

Notre nouveau jeu de données a 28 colonnes et 1586 lignes. Ce sont des données numériques à l'exception de la région du monde à laquelle appartient le pays et du régime politique. Et, nous avons 142 pays au total, pour 10 régions et 16 années sans valeur manquante. Ces nouvelles variables nous font donc perdre quelques lignes et quelques pays mais le jeu de données reste très important.

Nos variables de départ sont toujours distribuées de la même manière :



Voici la distribution de nos nouvelles variables :

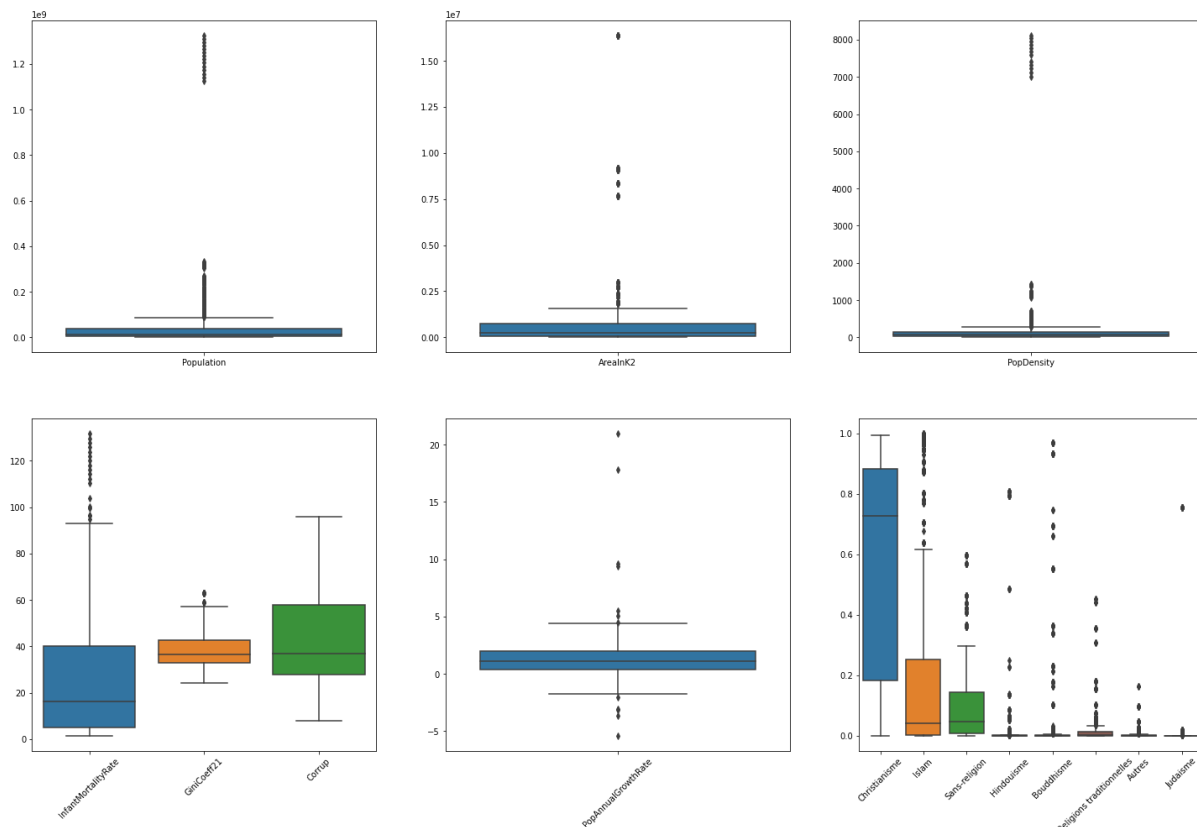
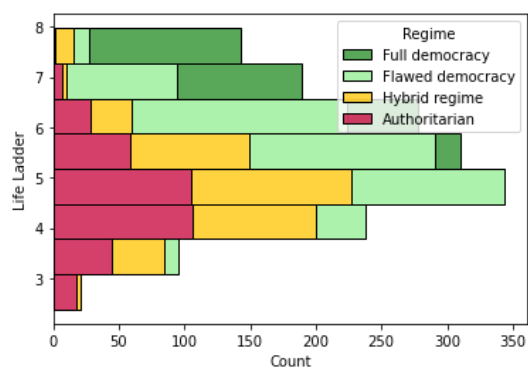


Image 15 : Boîte à moustache du nouveau jeu de données sur les nouvelles variables

Les échelles des nouvelles variables sont plus variées. La population, la superficie et la densité sont des variables avec de très grandes valeurs et souvent des valeurs extrêmes.

Concernant la distribution des religions, on note une forte dispersion pour certaines (ex: christianisme) mais des exceptions pour d'autres tels que l'Hindouisme, le Bouddhisme, les Religions traditionnelles, les Autres religions et le Judaïsme. Ces variables exceptionnelles seront à surveiller car elles peuvent être redondantes (elle expliquerait le pays par le pays ce qui a peu de sens pour nous).

Le reste des variables ont une dispersion sans de vrai valeurs exceptionnelles.



A propos, du régime politique, notre intuition semble bonne puisque la démocratie est visiblement un vecteur de bonheur, à l'opposé du régime autoritaire. Bien que cela semble relativement subtile si on regarde le graphique plus en détail. Mais c'est une variable qui peut mettre autrement en évidence les notions d'inégalités, de support social et de liberté.

Image 16 : Graphique nombre de pays avec tel régime par plage de life ladder

Étudiions ensuite une nouvelle matrice de corrélation :

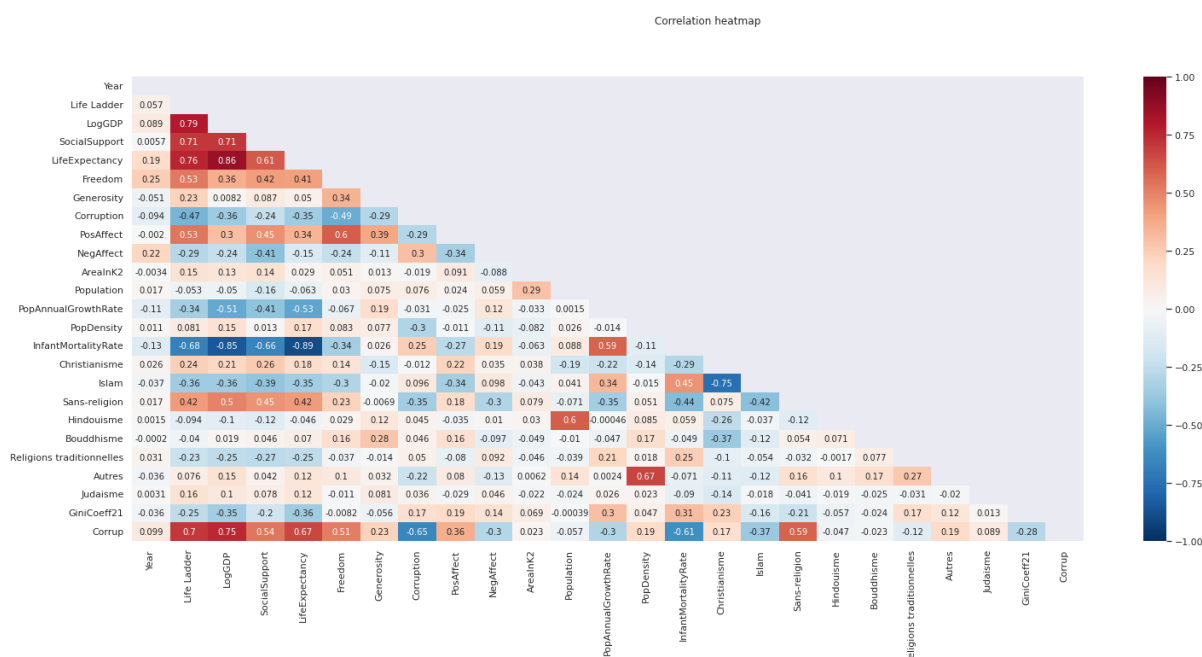


Image 17 : Matrice de corrélation du nouveau jeu de données

Désormais, les variables les plus corrélées au score de bonheur sont, de façon positive :

LogGDP : 0.79, LifeExpectancy : 0.76, SocialSupport : 0.71, la corruption mesurée : 0.7, Freedom : 0.53, PosAffect : 0.53, et la proportion de “sans religion” : 0.42.

et de façon négative : La mortalité infantile : -0.68, Corruption : -0.47, l’islam : -0.36

et la croissance de la population : -0.34.

Les autres variables sont moins corrélées, certaines plus ou moins négligeables.

Des variables ont aussi des corrélations fortes entre elles au delà du score de bonheur :

LogGDP : 0.86 avec LifeExpectancy, 0.75 avec la corruption mesurée, 0.71 avec SocialSupport et -0.85 avec le taux de mortalité infantile.

SocialSupport : 0.61 avec LifeExpectancy et -0.66 avec le taux de mortalité infantile.

LifeExpectancy : 0.66 avec la corruption mesurée et -0.89 avec le taux de mortalité infantile.

Freedom : 0.6 avec PosAffect

La corruption ressentie : -0.65 avec la corruption mesurée.

La population : 0.6 avec l’Hindouisme.

La densité de population : 0.67 avec les religions autres.

La mortalité infantile : -0.61 avec la corruption mesurée.

Le christianisme : -0.75 avec l’islam.

Certaines corrélations inter-variables sont simples à expliquer, d’autres le sont moins.

Avec ces données, nous allons donc tester plusieurs modèles pour tenter d’expliquer le bonheur.

Démarche et sélection des variables

Régression Linéaire multiple

Fort de la première analyse des variables fournies et de leurs corrélations, nous avons décidé dans un premier temps d'appliquer un modèle de régression linéaire simple. Il s'agit de créer une fonction linéaire qui permet de prédire une variable en fonction d'autres variables sur la base d'observations passées. Cette simplification d'un nuage d'observations en une droite nous permet d'associer un poids à chaque variable explicative sur la variable expliquée, et surtout nous donne la possibilité de faire des prédictions de la variable cible si tant est qu'on ait les données des variables explicatives.

Ce premier modèle de régression linéaire simple appliqué sur nos données a donné un score de 79%, c'est à dire que 79% de la variation du score de bonheur observé peut être expliqué par les variables explicatives. Les coefficients de régression retenus pour chaque variable sont dans le tableau suivant. Et le graphique à la droite du tableau montre l'étendue du taux d'erreur, soit la différence entre le taux de bonheur prédit et celui observé pour une partie du jeu de données.

| Variable | Coeff |
|--------------------------|--------|
| Sentiments positifs | 1.990 |
| Support Social | 1.805 |
| Sentiment de liberté | 0.488 |
| Sentiment de générosité | 0.438 |
| Sentiments négatifs | 0.417 |
| PIB | 0.392 |
| Espérance de vie | 0.025 |
| Année | -0.011 |
| Perception de corruption | -0.741 |

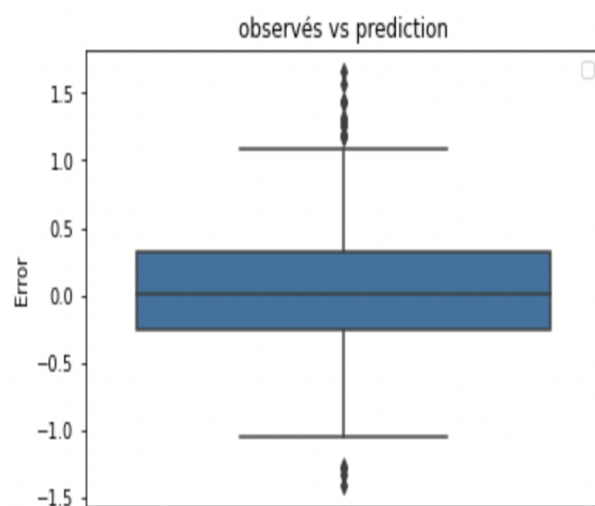


Tableau 1 : Coefficient de régression retenus pour chaque variables

Image 18 : Dispersion de la précision de prédiction

On observe que le taux d'erreur de prédiction du score de bonheur est, pour la grande majorité des pays, inférieur à 1 et pour la moitié d'entre eux inférieur à 0.5. Ce qui signifie que les prédictions ne sont pas très éloignées de la réalité et restent donc dans la même catégorie. Autrement dit qu'un pays ayant un score de bonheur à 7 ou 8 peut être considéré comme un très bon pays tandis qu'un autre ayant un score à 2 ou 3 est considéré comme très mauvais.

Pas de grande surprise concernant les coefficients associés à chaque variable, ni à propos de leur poids respectifs.

Cette première régression nous a amené à observer si l'on pouvait représenter la répartition des pays selon leur taux de bonheur et s'ils étaient répartis en catégories.

Nous sommes parvenus au graphique suivant avec tous les pays classés chaque année :



15

Analysons maintenant la représentation des variables selon les 2 axes retenu par l'ACP :

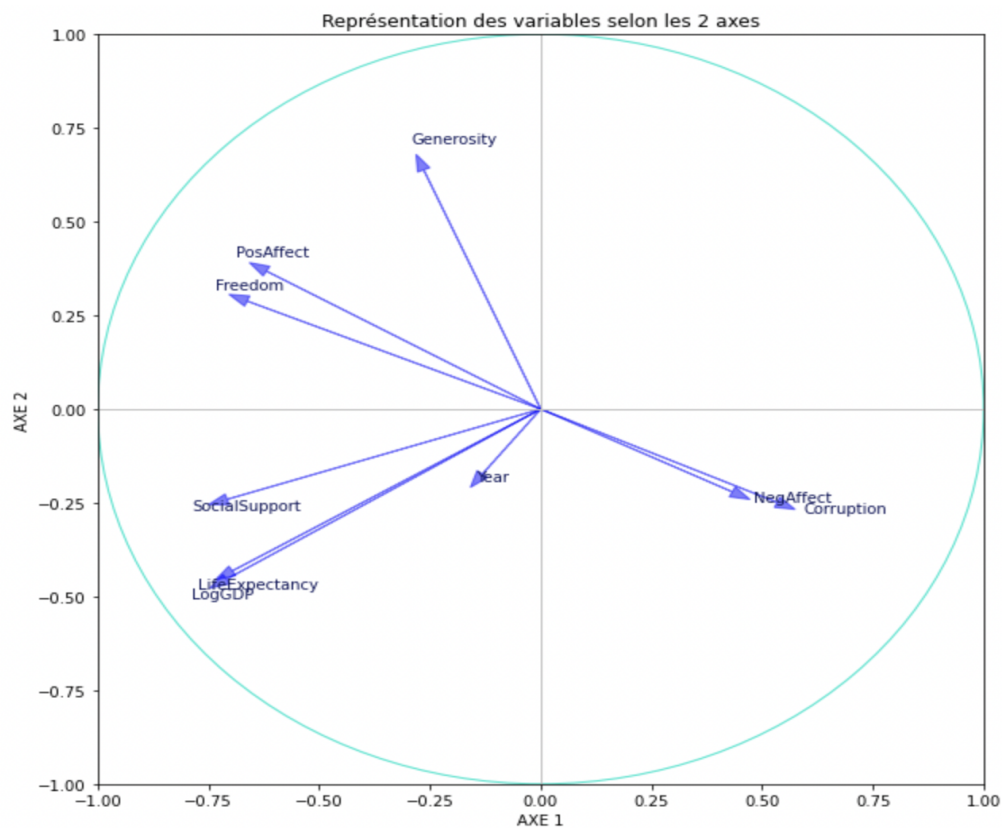


Image 20 : Analyse des variables en fonction des 2 axes ACP

On remarque une dépendance entre le PIB par tête (LogGDP) et l'espérance de vie (LifeExpectancy), de même qu'entre la perception de la corruption (Corruption) et le niveau de sentiment négatif (NegAffect). On peut donc se poser la question d'en garder une sur les 2 à chaque fois. La même question se pose pour le sentiment de liberté (Freedom) et le niveau de sentiments positifs (PosAffect). Bien que ces variables reflètent des aspects sensiblement différents.

De plus, on peut observer qu'aucune variable initialement présente n'a d'influence vers le cadran en haut à droite. Or dans le schéma précédent, on observe que beaucoup des pays ayant un faible taux de bonheur sont représentés dans ce cadran. On peut en conclure que les variables initialement fournies ne sont pas suffisantes pour expliquer le taux de bonheur.

Modèle de classification

Suite à la constatation que les pays peuvent être classés en catégories, nous avons décidé d'étudier un modèle de classification. Ici, l'arbre de décision a été préféré pour sa simplicité.

Nous avons appliqué le modèle d'abord en ayant un taux de bonheur en 2 classes (séparées par la médiane) ; et d'autre part en faisant une échelle de 0 à 10 en arrondissant les valeurs données. Les scores (taux de bonnes prédictions) obtenues sont de : 90% lorsque nous avons 2 classes et de 61% lorsque les valeurs sont arrondies. Le score de 90% indique la facilité du modèle à diviser les pays en 2 catégories. Ce qui confirme bien l'observation faite à l'aide de la réduction de dimensions plus haut où l'on avait constaté la division claire en 2 des pays selon leur taux de bonheur. Lorsque les valeurs sont arrondies à l'unité (10 classes), les valeurs bien prédites ont une erreur inférieure à 0,5. Dans ce cas, nous avons un score de 61%. Or, lors de la régression linéaire vue précédemment, nous avons vu que la moitié des pays avaient un taux d'erreur inférieur à 0,5. Ce qui signifie que le modèle Arbre de Décision est meilleur que la régression linéaire.

| | Poids des variable pour 2 classes |
|----------------------|--------------------------------------|
| PIB | 0.586 |
| Espérance de vie | 0.179 |
| Sentiment de liberté | 0.088 |
| Sentiments positifs | 0.087 |
| Support Social | 0.029 |

Tableau 2 : Poids des variables avec 2 classes

| | Poids des variables pour un score arrondi (10 classes) |
|--------------------------|---|
| Espérance de vie | 0.603 |
| PIB | 0.147 |
| Support Social | 0.132 |
| Sentiments positifs | 0.066 |
| Perception de Corruption | 0.033 |

Tableau 3: Poids des variables avec 10 classes

On observe qu'en fonction de la précision de la valeur que l'on souhaite pour le taux de bonheur, l'ordre d'importance des variables explicatives change. Ceci confirme le fait que le PIB par tête n'explique pas tout. Plus on veut être précis dans la détermination du taux de bonheur, moins le PIB a d'effet. En outre, plus l'on veut être précis, plus nous avons des valeurs négatives, comme la corruption avec une influence qui augmente.

Affichons la matrice de corrélation entre les valeurs réelles et celles déduite du modèle pour une partie du jeu de données réservé à l'évaluation de la qualité du modèle :

| score réel ↓ / score prédit → | 4 | 5 | 6 | 7 |
|-------------------------------|----|----|----|----|
| 3 | 10 | 4 | 0 | 0 |
| 4 | 33 | 25 | 2 | 0 |
| 5 | 15 | 66 | 26 | 0 |
| 6 | 0 | 13 | 78 | 9 |
| 7 | 0 | 2 | 19 | 32 |
| 8 | 0 | 0 | 0 | 8 |

Les taux de bonheur les mieux prédits sont ceux entre 5 et 7, des scores autour ou au-dessus de la médiane. Les extrêmes sont moins bien prédits. Pour les meilleurs, c'est dû au fait qu'il n'y en a pas beaucoup. Par contre, pour ceux étant en dessous de la médiane, cela confirme que les variables sont insuffisantes. Aussi, il peut y avoir un biais comme pour la corruption qui est très subjective puisque se basant sur un ressenti lié au climat politique, ainsi qu'au niveau d'information des sondés.

Tableau 4 : Matrice de prédiction (score réel vs score prédit)

On observe que les taux de bonheur les mieux prédits sont ceux entre 5 et 7, il s'agit de ceux autour ou au-dessus de la médiane. Les extrêmes sont moins bien prédits. Pour les meilleurs, c'est dû au fait qu'il n'y en a pas beaucoup. Par contre pour ceux étant en dessous de la médiane, cela confirme que les variables actuelles n'expliquent pas tout. Aussi, il peut y avoir un biais comme pour la corruption qui est très subjective puisque se basant sur un ressenti lié au climat politique, ainsi qu'au niveau d'information des sondés.

La spiritualité ou la religion peuvent également avoir un impact, tout comme le climat politique ou la démographie. Il serait intéressant également d'expliquer les cas comme l'Amérique latine qui a des variables matérielles non élevées mais qui par contre a un taux de bonheur parmi les plus élevés. Il y a également des contrastes inter régions à mettre en exergue éventuellement comme ce peut être le cas en Europe Occidentale entre les pays du Nord slaves et les pays du Sud latins.

Ajout de variables complémentaires

Dans un premier temps nous avons constaté que la perception de la corruption des sondés était très subjective et dépendante du climat politique perçu. Nous avons donc décidé d'ajouter un indicateur fourni par l'ONG Transparency International. ONG qui publie chaque année un indice de perception de la corruption qui se fonde sur un corpus d'indicateurs et de données, provenant notamment de la banque africaine de développement, de Freedom House, ou encore de la banque mondiale. Un score proche de 0 correspond à beaucoup de corruption, un score proche de 100 à peu de corruption. A noter que cet indice ne retient que la corruption dans le secteur public.

Ensuite, afin de limiter l'influence des variables matérielles nous avons décidé d'ajouter la notion de religion. Nous avons trouvé sur Wikipedia la distribution des religions dans chaque pays. A cela nous avons ajouté la notion de régime politique pour chaque pays grâce à un rapport de The Economist. A savoir, qu'il y a 4 types de régimes qui vont de autoritaire à démocratie (parfaite).

Enfin nous avons ajouté un indice d'inégalités, le coefficient de Gini, dans chaque pays, ainsi que des données démographiques que sont le taux de mortalité infantile, le taux de croissance et la densité de population à l'aide d'une institution américaine.

Modélisation avec les variables ajoutées

Nous avons testé à nouveau un arbre de décision avec ces nouvelles variables, et les scores obtenus (taux de bonnes prédictions) passent de 90% à 87% lorsque l'on souhaite positionner un pays entre 2 classes. Et de 61% à 59% lorsque l'on veut déterminer les valeurs arrondies (soit avec une erreur de 0.5 maximum sur l'échelle de 10).

Regardons quelles sont les 8 variables les plus importantes retenues par le modèle :

| Poids des variable pour 2 classes | |
|-----------------------------------|----------|
| Espérance de vie | 0.594 |
| Support Social | 0.136 |
| PIB | 0.130 |
| Taux de croissance (population) | 0.052 |
| Sans-religion | 0.040 |
| Islam | 0.016 |
| Sentiments négatifs | 0.014 |
| Sentiments positifs | 0.012575 |

Tableau 5 : Poids des variables avec 2 classes

| Poids des variables pour un score arrondi (10 classes) | |
|--|----------|
| Espérance de vie | 0.448 |
| PIB | 0.288 |
| Support Social | 0.103 |
| Taux de croissance (population) | 0.049 |
| Population | 0.043 |
| Sentiments positifs | 0.029 |
| Corruption ajoutée | 0.026 |
| Taux d'inégalité | 0.014765 |

Tableau 6 : Poids des variables avec 10 classes

On peut observer que les nouvelles variables ont une influence sur le modèle, une influence certes modeste mais non négligeable, et surtout intéressante à interpréter.

Suite à cela nous avons testé le modèle de classification en forêts aléatoires. Nous avons constaté que ce modèle a tendance à faire du sur apprentissage et à apprendre par cœur le jeu de données réservé à l'entraînement du modèle. Cependant, il reste intéressant de regarder l'ordre d'importance des variables dans la détermination du taux de bonheur :

| Importance des variables pour déterminer la classe | |
|--|-------|
| Espérance de vie | 0.095 |
| PIB | 0.092 |
| Support Social | 0.080 |
| Corruption ajoutée | 0.060 |
| Taux mortalité infantile | 0.059 |
| Taux de croissance (population) | 0.055 |
| Sentiments positifs | 0.054 |
| Sentiment de liberté | 0.046 |
| Perception de Corruption | 0.046 |
| Population | 0.043 |

Tableau 7 : Importance des variables

On observe que les variables ajoutées ont une influence non négligeable sur la détermination du taux de bonheur. C'est notamment le cas de la corruption dans la fonction publique ainsi que des données démographiques.

Enfin, nous avons testé ce classificateur, non plus sur l'ensemble du monde, mais seulement sur les pays d'Amérique latine et Caraïbes et ensuite sur ceux d'Europe Occidentale.

Sur l'Amérique latine et Caraïbes nous l'ordre d'influence des variables suivant :

| Importance des variables pour déterminer la classe | |
|--|-------|
| PIB | 0.124 |
| Espérance de vie | 0.078 |
| Sentiments positifs | 0.070 |
| Taux de mortalité infantile | 0.066 |
| Sentiment de liberté | 0.061 |
| Sentiments négatifs | 0.061 |
| Densité de population | 0.060 |
| Support Social | 0.059 |

Tableau 8 : Importance des variables pour l'Amérique latine

On remarque que l'on retrouve comme au niveau mondial l'influence des données démographiques, même si ici au lieu d'avoir le taux de croissance de la population c'est plutôt la densité de la population qui est particulièrement importante.

Quant à l'Europe Occidentale, nous avons :

| Importance des variables pour déterminer la classe | |
|--|-------|
| Perception de Corruption | 0.135 |
| Corruption secteur public | 0.084 |
| Espérance de vie | 0.075 |
| Sentiments négatifs | 0.072 |
| PIB | 0.071 |
| Sentiment de liberté | 0.071 |
| Support Social | 0.057 |
| Population | 0.044 |

Tableau 9 : Importance des variables pour l'Europe

Cette fois c'est la perception de la corruption par les sondés, et la corruption dans les institutions publiques qui ont la plus grosse influence. Elle est même supérieure aux variables matérielles comme le PIB. En Europe Occidentale il y a une séparation en deux entre les pays d'Europe du Nord (notamment scandinaves) qui eux sont réputés pour la transparence de la politique, et d'un autre côté les pays du Sud latins. Dans ces derniers, le niveau de corruption des classes dirigeantes reste un sujet récurrent.

Dans cette partie nous avons testé différents modèles. Nous avons vu d'une part que le modèle de régression linéaire avait de bonnes performances mais il reste limité dans pour interpréter les poids des variables dans la détermination du score de bonheur.

L'arbre de décision, quant à lui, à de meilleures performances et nous donne les variables qu'il a retenu par ordre d'importance pour déterminer le taux de bonheur. Ceci nous a permis de constater les limites des variables fournies initialement, et ensuite la pertinence de l'ajout d'autres variables pouvant expliquer le taux de bonheur. C'est notamment le cas des données démographiques et des données concernant la corruption des institutions publiques.

Ces constatations sont confirmées grâce au modèle des forêts aléatoires qui, lui, apprend le jeu de données par cœur. Nos observations ont permis de mettre en valeur des différences entre régions : en fonction des régions le taux de bonheur peut être expliqué par des variables différentes. Le modèle de forêts aléatoires a de meilleures performances, cependant l'apprentissage par cœur du jeu de données apporte un biais aux futures prédictions.

Ces toutes ces raisons qui nous ont fait retenir un modèle de régression linéaire régularisée, dit Lasso, qui sélectionne les variables ayant le plus grand effet sur l'estimation du score de bonheur. Ce modèle nous a donné des performances très encourageantes. C'est pourquoi nous lui consacrons les interprétations et études les plus détaillées. A noter que nous avons également ajouté cette fois une nouvelle série de ligne au jeu de données : les notions de dystopie et d'utopie qui possèdent chaque année, respectivement, les pires et les meilleurs scores à toutes les réponses du sondage World Happiness Report. De sorte à aider le modèle à prédire les scores les plus extrêmes.

Modèle retenu et interprétation

Modèle de Lasso

A l'aide des démarches précédentes, nous avons retenu un modèle de régression linéaire dit "de Lasso" qui nous permet de tester plusieurs configurations à la marge. En effet, il nous semblait important de pouvoir analyser plusieurs combinaisons des variables retenues pour comparer les comportements du modèle retenu avec un modèle alternatif.

Le méthode de Lasso a pour particularité d'affecter des coefficients aux variables explicatives de telle sorte qu'une variable ayant un faible poids sur la variable expliquée soit simplement annulée. La méthode fonctionne avec un paramètre alpha représentant la sensibilité de la méthode. Dans l'optique de maximiser nos performances de prédictions, nous avons également retenu la méthode de validation croisée permettant de retenir le paramètre alpha ayant les meilleurs résultats. Ainsi nous obtenons deux modèles qui expliquent plus de 80% de la variance observée à la fois sur un ensemble d'entraînement et sur un ensemble de test.

Comparons deux séries de variables différentes avec la même méthode de Lasso, voici les coefficients les plus importants associés à chaque variable explicative du score de bonheur :

| Variable | Coeff | Variable | Coeff |
|--|--------------|--|-------------|
| Religion - Judaïsme | +1,38 | Religion - Judaïsme | 1,69 |
| Sondage - Support social | +1,35 | Région - Amérique latine et caraïbes | 0,96 |
| Sondage - Sentiments positifs | +1,26 | Région - Amérique du nord et Australie | 0,73 |
| Sondage - Sentiment de liberté | +0,79 | Economie - PIB par tête (logarithme) | 0,51 |
| Région - Amérique latine et caraïbes | +0,53 | Région - Europe de l'Ouest | 0,46 |
| Sondage - Sentiment de générosité | +0,45 | Région - Asie du Sud-Est | 0,45 |
| Economie - PIB par tête (logarithme) | +0,42 | Religion - Islam | 0,33 |
| Région - Amérique du nord et Australie | +0,41 | Régime politique - Démocratie | 0,22 |
| Région - Europe de l'Ouest | +0,30 | Religion - Christianisme | 0,08 |
| Religion - Islam | +0,12 | Religion - Hindouisme | 0,04 |
| Régime politique - Démocratie | +0,07 | Région - Afrique sub-saharienne | 0,02 |
| Régime politique - Dictature | -0,05 | Economie - Espérance de vie | 0,01 |
| Religion - Hindouisme | -0,08 | Régime politique - Dictature | -0,04 |
| Région - Russie et ex-URSS | -0,08 | Région - Asie de l'Est | -0,10 |
| Religion - Sans-religion | -0,10 | Régime politique - Démocratie imparfaite | -0,15 |
| Région - Moyen-orient et Afrique du Nord | -0,12 | Région - Moyen-orient et Afrique du Nord | -0,33 |
| Régime politique - Démocratie imparfaite | -0,21 | | |
| Région - Afrique sub-saharienne | -0,28 | | |
| Sondage - Sentiment négatifs | -0,35 | | |
| Religion - Bouddhisme | -0,49 | | |

Tableau 10 : Coefficient des variables avec Lasso

La différence entre les deux séries reposent sur l'absence des autres réponses au sondage qui sont fortement corrélées au score de bonheur déclaré en même temps. Il s'agit de constater quelles variables absorbent le poids de ces questions et donc expliquent les perceptions des sondés. On remarquera que le paramètre alpha retenu est fin et conserve la majorité des variables, nous laissant interpréter quelques aspects moins importants mais significatifs.

Les réponses annexes au sondage ont donc un poids important, les sentiments positifs en particulier reposent sur une question très proche du score de bonheur. Naturellement, le sentiment de pouvoir compter sur les autres, d'être libre de faire des choix et enfin la perception de la générosité sont des éléments qui jouent positivement sur le bonheur. Pas de surprise concernant les sentiments négatifs.

C'est à ce stade que l'étude devient plus intéressante puisque l'on observe que le PIB, toujours important dans le modèle, cède sa première place à des régions voire à des religions que nous décrirons ensuite. L'espérance de vie est presque inexistante dans ce modèle mais nous savons qu'elle possède une très forte corrélation avec le PIB et avec les régions du monde. Le poids des richesses est donc considérable mais nous verrons également comment l'interpréter autrement.

Les régions ont un fort pouvoir d'explication du bonheur car elles absorbent plusieurs aspects que nous n'avons pas détaillé : antécédents historiques, climat, unions et tensions locales. Les pays "du Nord" restent mieux lotis que les pays "du Sud". C'était notre intuition, les religions sont partiellement localisées mais restent des facteurs importants par rapport à la perception du bonheur. Il convient de remarquer que l'Israël est le seul pays où le judaïsme est majoritaire et de très loin, or c'est un pays extrêmement riche, c'est la raison pour laquelle le poids est si positif. Il en va de même pour l'hindouisme très concentré en Inde où la pauvreté atteint des niveaux extrêmes. Ce qui est particulièrement intéressant, c'est à quel point le retrait des questions du sondage propre à la perception de la vie favorise l'ensemble des religions, l'islam, le christianisme et même l'hindouisme qui devient un facteur positif. Il est donc indéniable que la spiritualité est synonyme de bien-être, pour preuve on retrouve l'absence de religion dans les facteurs négatifs. Enfin, l'Afrique sub-saharienne remonte à flot sans les questions du sondage, on peut donc en conclure que cette région est plus optimiste malgré les contraintes locales.

Concernant les régimes politiques, on observe que la démocratie est positive, et probablement associée à la notion de liberté alors que la dictature est bien entendu négative. On constate l'absence de l'indice de corruption qui peut disparaître par le biais des régimes politiques. Plus étonnant cependant, le facteur le plus négatif est la démocratie imparfaite dans le 1er modèle (tandis que c'est un facteur positif mais très faible dans le second modèle : voir annexes). Il est probable que ce soit un facteur important dans les pays riches où l'on se sent libre pour autant (la sensation de liberté étant absente du second modèle). Et dans le second modèle c'est donc la démocratie parfaite qui a le plus de poids car elle peut valoir liberté, support social ou encore sentiments positifs. Enfin, on constate l'absence de poids des inégalités et de la densité de population qui est certainement absorbée par les autres variables économiques.

Dans l'absolu aucun des modèles n'est meilleur d'un point de vue réel, tous deux mettent en évidence la complexité d'expliquer le bonheur par des variables corrélées entre elles.

Modèle à inertie

Durant l'analyse des données nous avons observé que l'année jouait peu dans le score de bonheur obtenu dans chaque pays, pourtant l'objectif de ce projet de l'ONU est bien de favoriser le développement de cet indicateur et donc la dimension temporelle est fondamentale. Avec cette idée en tête nous avons considéré une hypothèse forte qui consiste à dire que le score de bonheur a une très forte inertie comme l'ensemble des variables du modèle car le poids de l'histoire d'un pays est indissociable de l'explication du bonheur que l'on peut en faire.

C'est donc ainsi que nous avons testé une nouvelle variable supplémentaire, le score de l'année passée. Bien entendu nous savions que cela donnerait de très bonnes prédictions (score R2 de 90) pour les raisons évoquées auparavant. Mais justement ce qui nous intéressait était de savoir à quel point l'inertie est forte, et à l'inverse quelles sont les variables qui pèsent toujours dans l'explication du score de bonheur. Les autres questions du sondage ont été retirées comme précédemment.

| Variable | Coeff | Variable | Coeff |
|---|-------------|--|--------------|
| Inertie - Score de bonheur n-1 | 0,72 | Religion - Christianisme | |
| Religion - Judaïsme | 0,43 | Région - Afrique sub-saharienne | |
| Région - Amérique latine et caraïbes | 0,24 | Religion - Sans-religion | |
| Economie - PIB par tête (logarithme) | 0,14 | Région - Europe de l'Est et centrale | |
| Région - Amérique du nord et Australie | 0,13 | Religion - Bouddhisme | |
| Région - Europe de l'Ouest | 0,09 | Région - Asie de l'Est | |
| Religion - Islam | 0,07 | Religions traditionnelles | |
| Régime politique - Démocratie | 0,06 | Religion - Autres religions | |
| Région - Asie du Sud-Est | 0,05 | Economie - Densité de population | |
| Région - Russie et ex-URSS | 0,03 | Economie - Coefficient de Gini (indice d'inégalités) | -0,001 |
| Régime politique - Hybride | 0,006 | Régime politique - Dictature | -0,005 |
| Economie - Espérance de vie | 0,005 | Régime politique - Démocratie imparfaite | -0,04 |
| Economie - Corruption (indice) | 0,003 | Région - Asie du Sud | -0,05 |
| Religion - Hindouisme | | Région - Moyen-orient et Afrique du Nord | -0,16 |

Tableau 11 : Coefficient modèle inertie

Ce qui est particulièrement intéressant, c'est dans quelle mesure le PIB a considérablement réduit son influence sur le bonheur quand on considère l'inertie des années passées. C'était l'objectif de cette série de variables qui peut mettre en évidence les autres composantes du bonheur.

On retrouve le judaïsme associé à l'État d'Israël qui est trop singulier pour être expliqué par notre modèle. En revanche, on peut conclure que l'Amérique latine a une certaine capacité à expliquer le score de bonheur qui n'est pas lié au poids de l'histoire mais à quelque chose qui pourrait s'assimiler à une certaine culture du bonheur. A l'inverse, la région moyen-orient et Afrique du Nord concentre le poids négatif sur l'échelle de bonheur, ce qu'on peut associer à une longue instabilité avec des conflits armés, guerres civiles ou internationales. Ce modèle ne dit donc pas beaucoup plus qu'il confirme le postulat lié à l'inertie du score de bonheur, et à celle des richesses.

Capacité de prédiction

Enfin, nous souhaitons exploiter notre modèle pour prédire les scores de 2021 mais le problème est que nos variables explicatives ne sont pas disponibles et les estimer fausserait totalement la démarche. Nous avons donc choisi de prédire les scores de 2020 en retirant ces données de l'ensemble d'entraînement (ce modèle obtient un score R2 de 77 puisqu'il bénéficie d'un entraînement sur tous les pays mais pas des résultats du sondage).

| Variable | Coeff | Variable | Coeff |
|--|-------|--|--------|
| Religion - Judaïsme | 1,91 | Régime politique - Dictature | |
| Région - Amérique latine et caraïbes | 1,01 | Région - Asie du Sud | |
| Région - Amérique du nord et Australie | 0,80 | Région - Europe de l'Est et centrale | |
| Région - Asie du Sud-Est | 0,51 | Région - Russie et ex-URSS | |
| Economie - PIB par tête (logarithme) | 0,50 | Religion - Autres religions | |
| Région - Europe de l'Ouest | 0,49 | Régime politique - Hybride | |
| Religion - Islam | 0,30 | Economie - Densité de population | |
| Régime politique - Démocratie | 0,24 | Economie - Coefficient de Gini (indice d'inégalités) | -0,008 |
| Religion - Hindouisme | 0,13 | Religion - Bouddhisme | -0,04 |
| Religion - Christianisme | 0,08 | Région - Asie de l'Est | -0,05 |
| Région - Afrique sub-saharienne | 0,03 | Religion - Sans-religion | -0,07 |
| Economie - Espérance de vie | 0,010 | Religion - Religions traditionnelles | -0,11 |
| Economie - Corruption (indice) | 0,009 | Régime politique - Démocratie imparfaite | -0,17 |
| Economie - Corruption (indice) | 0,009 | Région - Moyen-orient et Afrique du Nord | -0,28 |

Tableau 12 : Coefficient des variables du modèle choisi

On retrouve de grandes similarités avec les observations précédentes.

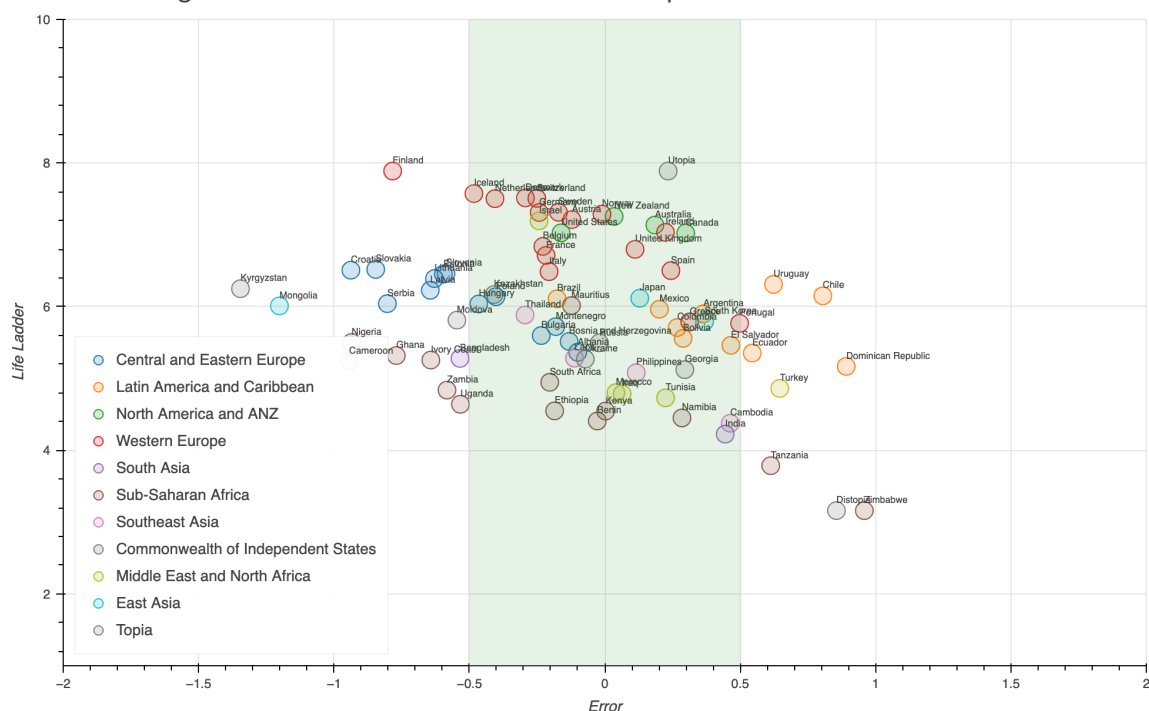


Image 21 : Score life ladder prédit et erreur de prédiction par pays

Ce graphique nous permet de constater que notre capacité à prédire permet de faire relativement peu d'erreurs au-delà de 0,5 pts sur 10 (inclus dans la bande verte). Les erreurs semblent être plus concentrées sur les régions Afrique sub-saharienne, Amérique latine et Europe centrale. Dans un premier temps on constate que nous sous-estimons un peu plus les états avec des scores relativement élevés, et nous sous-estimons un peu plus les états avec des scores relativement faibles. Notre modèle tend donc vers un score relativement moyen et détecte mal des exceptions.

Au-delà de ce constat il est difficile de tracer une conjecture par rapport à ce qu'il s'est réellement passé en 2020 dans l'esprit des sondés, que ce soit du point de vue court terme et long terme. Mais dans un premier temps on peut voir que nous

Pays dont le score a été particulièrement sur-estimé par le modèle :

- République dominicaine, Chili, Uruguay, Equateur
- Zimbabwe, Tanzanie
- Turquie

Nous constatons que l'Amérique latine était un fort facteur du bonheur dans le modèle et donc il est naturellement aussi responsable d'une partie des erreurs de prédiction. La seule chose que nous pouvons dire est que ces pays se distinguent particulièrement des autres pays d'Amérique latine de manière générale, ou particulièrement en 2020. Aussi il est possible que les autres pays de la région pèsent plus lourd dans les statistiques notamment avec un PIB par tête bien plus élevé et avec un sondage plus récurrent que les pays précédents.

Pays dont le score a été particulièrement sous-estimé par le modèle :

- Croatie, Slovaquie, Serbie, Lituanie, Lettonie, Slovenie, Moldavie
- Nigeria, Cameroun, Ghana, Côte d'Ivoire, Zambie, Ouganda
- Kirghizistan, Mongolie, Bangladesh
- Finlande

Les pays d'Europe centrale regroupent de nombreux pays de l'ex-URSS dont les pays des Balkans qui ont connu des conflits militaires bien plus récents que le reste de l'Europe.

Enfin, à propos des pays d'Afrique sub-saharienne, il semble important de noter que la région est très vaste et regroupe des pays avec des différences considérables en termes d'économie, d'histoire et de climat avec des valeurs extrêmes dans nos variables. Notre modèle manque donc de précision au niveau de cette région du monde.

Ces résultats sont également intéressants parce qu'ils nous donnent peut-être celui qu'on aurait pu observer en l'absence de la pandémie que nous avons connu. Et dans le même temps, cela nous rappelle que les prédictions reposant sur l'inférence statistique dépendent de la stabilité du paradigme dans lequel nous vivons.

Conclusion

Avec une modélisation simple, il est facile d'expliquer l'essentiel du score de bonheur avec le PIB par tête qui nous permet de tirer une conclusion rapide, l'argent fait le bonheur. Mais il semble bien évident que ce n'est pas suffisant voire trompeur, car cette conclusion réside dans l'histoire moderne qui a concentré les richesses dans les pays dit "du Nord" quasi exclusivement au détriment des pays dit "du Sud" avec en prime des conséquences climatologiques et sociales dramatiques sur l'ensemble du monde.

Notre démarche analytique nous a permis d'établir un modèle plus complexe qui réduit le poids du PIB au profit de composantes sociales, politiques et religieuses qui semblent cette fois essentielles dans la perception du bonheur. La liberté, la fraternité, la générosité, la spiritualité s'opposent naturellement à la corruption, à la tyrannie et aux inégalités; et nous pouvons effectivement le démontrer à l'aide de statistiques.

Bien sûr, prétendre savoir comprendre et prédire le bonheur de l'humanité était une cible totalement hors de portée de nos outils, au même titre qu'il semble illusoire de pouvoir écrire un algorithme du bonheur universel reposant sur des outils cartésiens.

Pour autant, aussi critiquable qu'elle soit, cette approche mathématique nous permet de mettre en évidence des caractères du bonheur déclaré qui vont dans le sens d'une vision moins matérialiste et plus sociale du bonheur, à contre-courant d'un capitalisme exponentielle en bout de course dans un monde de ressources finies. Pour autant, cette démarche analytique n'est pas vaine pour aider le progrès général, et notre approche pourrait être considérablement améliorée avec des données plus précises, plus pertinentes sur les mêmes aspects et dans d'autres domaines.

La notion de religion notamment devrait être plus approfondie du point de vue des pratiquants, de la nature des croyances et de leurs préceptes fondamentaux. Au-delà des religions, la nature des sociétés et de leur histoire devrait révéler des caractères plus ou moins individualistes de la perception du bonheur.

Des évènements locaux ou internationaux pourraient être agrégé pour tenir compte de catastrophes ou à l'inverse de ferveur nationale propre à une simple victoire sportive par exemple. Les richesses sont sans aucune doute loin d'être décorés du bonheur perçu et donc il faudrait approfondir les notions de distribution nationale, de ressources locales ou d'exploitation étrangère.

Enfin, il sera probablement judicieux de retenir la notion d'inertie évoquée dans ce rapport, en tenant compte du poids de l'histoire qui à des conséquences sur des générations : conflits, invasions, alignement idéologique, unité nationale, etc.

A peine ouverte, la HapPy Factory ferme déjà ses portes, trop vite concurrencée par la première case du calendrier de l'avent et l'approche des fêtes de fin d'année.

Sources

Données du World Happiness Report (Gallup World Poll) 2015/2020

Rapport mondial sur le bonheur (en anglais : World Happiness Report) propose une mesure du bonheur publiée par le Réseau des solutions pour le développement durable des Nations unies chaque année depuis 2012.

Proportion des religions et des croyances dans les pays du monde en 2010 - Pew Research Center

https://fr.wikipedia.org/wiki/Liste_des_pays_par_religion

Indice de perception de la corruption 2012/2020 - Transparency International

https://fr.wikipedia.org/wiki/Indice_de_perception_de_la_corruption

Les deux jeux de données précédents sont récoltées à l'aide d'une méthode de scrapping en python.

Indice de démocratie 2020 - The Economist

Indicateur évalué en fonction de 60 critères concernant le pluralisme, les libertés et la culture civique

Classification en type : Démocratie, démocratie imparfaite, hybride, dictature

https://en.wikipedia.org/wiki/Democracy_Index

Données démographiques - United States Census Bureau

<https://www.census.gov/>

Coefficient de Gini (indice d'inégalité) - 2020

<https://worldpopulationreview.com/country-rankings/wealth-inequality-by-country>

Annexes

Méthode de traitement religion

Les données concernant les proportions ont été ajoutées depuis un tableau depuis wikipedia. Elles ont été ajoutées via la méthode de webscrapping. Voici le code utilisé ci-dessous :


```
1) Définition d'une variable reprenant le lien de la page wikipedia
page_religion = "https://fr.wikipedia.org/wiki/Liste_des_pays_par_religion"

2) Récupération du tableau via le nom de la balise html (table) et la classe de celui-ci
table_class='wikitable sortable zebra jquery-tablesorter'
response=requests.get(page_religion)
print(response.status_code)
soup = BeautifulSoup(response.text, 'html.parser')
religion_table=soup.find('table', attrs={'class': 'wikitable sortable zebra'})

3) Lecture du tableau et transformation en DataFrame Pandas

religion=pd.read_html(str(relation_table))
religion=pd.DataFrame(relation[0])
```

Le tableau extrait de wikipedia est de la forme ci-contre :

| Pays | Christianisme | Islam | Sans-religion | Hindouisme | Bouddhisme | Religions traditionnelles | Autres | Judaïsme |
|---|---------------|-------|---------------|------------|------------|---------------------------|--------|----------|
|  Afghanistan | 0.1% | 99.7% | <0.1% | <0.1% | <0.1% | <0.1% | <0.1% | <0.1% |

Avant de pouvoir l'exploiter, il a donc fallu faire de la Data Quality via les étapes suivantes :

- 1) Suppression du caractère ‘%’
- 2) Remplacement du ‘<0.1’ par ‘0’
- 3) Positionnement des valeurs en type ‘float’
- 4) Division des nombres par 100
- 5) Traduction des pays en anglais car ils sont en français sur le DataFrame initial
- 6) Correction des noms de pays qui diffèrent des données initiales

Enfin le DataFrame contenant les religions a été joint au DataFrame initial.

Tableau complet des coefficient du modèle de Lasso initial

| Variable | Coeff | Variable | Coeff |
|--|--------|--|--------|
| Religion - Judaïsme | +1,38 | Religion - Judaïsme | +1,69 |
| Sondage - Support social | +1,35 | Région - Amérique latine et caraïbes | +0,96 |
| Sondage - Sentiments positifs | +1,26 | Région - Amérique du nord et Australie | +0,73 |
| Sondage - Sentiment de liberté | +0,79 | Economie - PIB par tête (logarithme) | +0,51 |
| Région - Amérique latine et caraïbes | +0,53 | Région - Europe de l'Ouest | +0,46 |
| Sondage - Sentiment de générosité | +0,45 | Région - Asie du Sud-Est | +0,45 |
| Economie - PIB par tête (logarithme) | +0,42 | Religion - Islam | +0,33 |
| Région - Amérique du nord et Australie | +0,41 | Régime politique - Démocratie | +0,22 |
| Région - Europe de l'Ouest | +0,30 | Religion - Christianisme | +0,08 |
| Religion - Islam | +0,12 | <i>Religion - Hindouisme</i> | +0,04 |
| Régime politique - Démocratie | +0,07 | <i>Région - Afrique sub-saharienne</i> | +0,02 |
| Régime politique - Hybride | +0,063 | Economie - Espérance de vie | +0,01 |
| Région - Asie du Sud-Est | +0,033 | Economie - Corruption (indice) | +0,009 |
| <i>Religion - Christianisme</i> | +0,007 | Régime politique - Démocratie imparfaite | +0,008 |
| Economie - Corruption (indice) | +0,005 | Religion - Bouddhisme | 0,000 |
| Economie - Espérance de vie | +0,001 | Religion - Religions traditionnelles | 0,000 |
| Economie - Densité de population | 0,000 | Religion - Autres religions | 0,000 |
| Région - Europe de l'Est et centrale | 0,000 | Région - Europe de l'Est et centrale | 0,000 |
| Religion - Autres religions | 0,000 | Région - Russie et ex-URSS | 0,000 |
| Région - Asie de l'Est | 0,000 | Région - Asie du Sud | 0,000 |
| Religion - Religions traditionnelles | 0,000 | Economie - Densité de population | 0,000 |
| Région - Asie du Sud | 0,000 | Religion - Sans-religion | -0,005 |
| Economie - Coefficient de Gini (indice d'inégalités) | -0,007 | Economie - Coefficient de Gini (indice d'inégalités) | -0,007 |
| Régime politique - Dictature | -0,05 | Régime politique - Dictature | -0,04 |
| <i>Religion - Hindouisme</i> | -0,08 | Région - Asie de l'Est | -0,10 |
| Région - Russie et ex-URSS | -0,08 | Régime politique - Démocratie imparfaite | -0,15 |
| Religion - Sans-religion | -0,10 | Région - Moyen-orient et Afrique du Nord | -0,33 |
| Région - Moyen-orient et Afrique du Nord | -0,12 | | |
| Régime politique - Démocratie imparfaite | -0,21 | | |
| <i>Région - Afrique sub-saharienne</i> | -0,28 | | |
| Sondage - Sentiment négatifs | -0,35 | | |
| Religion - Bouddhisme | -0,49 | | |