ANALYSE DU BIEN-ÊTRE SUR TERRE  
  


CANALS Charlotte, CUVELLIEZ Nicolas, DETROZ Jérémie

Soutenance de projet

Formation Data Analyst fev23\_bootcamp\_da

Table des matières

[I - Découverte du projet et des données 3](#_Toc131613315)

[1 - Contexte 3](#_Toc131613316)

[2 - Découverte des données 3](#_Toc131613317)

[3 - Première analyse des données 4](#_Toc131613318)

[II – Présélection des variables 6](#_Toc131613319)

[1 - Choix des colonnes à conserver ou supprimer 6](#_Toc131613320)

[2 – Remplacement des valeurs manquantes 8](#_Toc131613321)

[III - Visualisations et Statistiques 9](#_Toc131613322)

[IV – Finalisation du jeu de données 14](#_Toc131613323)

[V - Modélisation 15](#_Toc131613324)

[1 - Contexte et début de la modélisation 15](#_Toc131613325)

[2 - Entraînement et validation des modèles 17](#_Toc131613326)

[3 - Fin des itérations de tests et sélection du modèle 18](#_Toc131613327)

[4 - Évaluation du modèle sur le jeu de test 22](#_Toc131613328)

[VI – Application du modèle sur de nouvelles données 23](#_Toc131613329)

[VII - Conclusion 24](#_Toc131613330)

[1 - Conclusion métier 24](#_Toc131613331)

[2 - Retour d’expérience 25](#_Toc131613332)

[3 - Axes d’amélioration 25](#_Toc131613333)

[4 - Rappel des sources 26](#_Toc131613334)

[VIII - Annexes 26](#_Toc131613335)

# I - Découverte du projet et des données

## 1 - Contexte

World Happiness est un rapport annuel qui évalue le niveau de bonheur et de bien-être des populations de différents pays du monde au travers de variables telles que le PIB par habitant, la durée de vie, les relations sociales, etc …

Source : <https://worldhappiness.report/>

Dans ce projet, nous allons effectuer une analyse approfondie des données collectées par le World Happiness Report.

Nous présenterons les données à l’aide de visualisations bien pensées et tâcherons de déterminer les combinaisons de facteurs permettant d’expliquer pourquoi certains pays sont mieux classés que les autres.

Nous avons tous les trois choisi cette thématique, car elle traite d’un sujet d’actualité et des valeurs humaines, auxquelles nous sommes particulièrement sensibles.

Même si nous ne disposons pas d’expertise technique ou scientifique, nous pensons que ce sujet est enrichissant et nous permettra de mieux comprendre notre environnement tout comme le bonheur humain. Nous pensons aussi que cette analyse pourra nous aider dans nos relations professionnelles, à mieux tenir compte des autres et à maintenir des relations de qualité.

## 2 - Découverte des données

Le dataset est librement accessible au lien suivant :

<https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021>

et est composé de deux fichiers csv. Nous allons donc les importer et les analyser plus en détails.

Le projet World Happiness Report dispose de deux fichiers csv.

L’objectif est de mieux comprendre la structure et les données incluses dans les deux jeux de données : quels sont les colonnes et le nombre de lignes ? Quels sont les types de variables ? Y a-t-il des valeurs manquantes ou des doublons ?

Commençons par le dataframe df\_2021 : le dataframe df\_2021 contient 149 lignes et 20 colonnes, dont 2 de types 'object' et 18 de types 'float64'. Il n’y a pas de valeurs manquantes (NaN) ni de doublons.

Analysons ensuite le dataframe df\_2021 : le dataframe df\_global contient 1949 lignes et 11 colonnes: 1 de type 'object', 1 de type 'int64' et 9 de types 'float64'. Il n’y a pas de doublons, mais 373 valeurs manquantes.

## 3 - Première analyse des données

Les deux dataframes de départ contiennent 5 colonnes avec exactement les mêmes noms et les mêmes types :

* **Country name** : (Nom du pays)

Il s’agit des pays participants au questionnaire. La colonne contient donc des variables catégorielles (nom de pays).

* **Social support** : (Support sociable)

Il s’agit de la réponse moyenne nationale à la question suivante :

« Si vous étiez en difficulté, avez-vous des parents ou des amis sur lesquels vous pouvez compter pour vous aider chaque fois que vous en avez besoin, ou non ?" La colonne contient donc des variables quantitatives continues, qui représentent les moyennes de réponses allant de 0 et 10.

* **Freedom to make life choices** : (Liberté de faire des choix de vie)

Il s’agit de la réponse moyenne nationale à la question suivante :

« Êtes-vous satisfait ou insatisfait de votre liberté de choisir ce que vous faites de votre vie ? » La colonne contient donc des variables quantitatives continues, qui représentent les moyennes de réponses sur une échelle de 0 et 10.

* **Generosity** : (Générosité)

Il s’agit du résidu de la moyenne nationale en régression des réponses à la question suivante : « Avez-vous donné de l’argent à un organisme de bienfaisance au cours du mois dernier ? » sur le PIB par habitant.

La colonne contient donc des variables quantitatives (en dollars).

* **Perceptions of corruption** : (Perception de corruption)

Il s’agit de la mesure de la moyenne nationale des réponses aux questions suivantes : « La corruption est-elle répandue ou non dans l’ensemble du gouvernement » et « La corruption est-elle répandue ou non dans les entreprises ? ». La perception globale n’est que la moyenne des deux réponses 0 ou 1. La colonne contient donc des variables quantitatives continues, qui représentent les moyennes de réponses allant d’une échelle de 0 et 10.

De plus, les deux dataframes contiennent également des colonnes qui sont similaires.

* **‘Life Ladder'/'Ladder score'** : (Échelle de vie, Hapiness Score)

Il s’agit de la réponse moyenne nationale à la question suivante :

"Veuillez imaginer une échelle, avec des marches numérotées de 0 (la plus basse) à 10 (la plus haute). Le haut de l’échelle représente la meilleure vie possible et le bas de l’échelle représente la pire vie possible. Sur quelle marche de l’échelle diriez-vous que vous vous sentez personnellement ce jour ?"

La colonne contient donc des variables quantitatives continues, qui représentent les moyennes de réponses allant d’une échelle de 0 et 10.

* **'Log GDP per capita'/'Logged GDP per capita'** : (PIB par habitant, afficher en logarithme)

Il s’agit des statistiques du PIB par habitant en fonction du pouvoir d’achat, calculées sur les données de 2020. La colonne contient donc des variables quantitatives (en dollars).

* **'Healthy life expectancy at birth'/'Healthy life expectancy'** : (Espérance de vie en bonne santé à la naissance)

Les espérances de vie en bonne santé à la naissance sont basées sur les données extraites du rapport mondial de l’Organisation Mondiale de la Santé (OMS), disponibles pour les années 2000, 2005, 2010, 2015 et 2016.

La colonne contient donc des variables quantitatives (ages).

À première vue, **les colonnes communes nous paraissent être les variables clefs et les plus pertinentes pour atteindre notre objectif** : déterminer les combinaisons de facteurs permettant d’expliquer pourquoi certains pays sont mieux classés que les autres dans le World Happiness Report.

De plus, **la variable cible est ici : 'Life Ladder'/'Ladder score'** qui correspond à un "Happiness score", une échelle de vie qui estime le bonheur des pays inclus dans le World Happiness Report.

Enfin, avant de clôturer cette première analyse, nous tenions à noter que les deux jeux contiennent des données complexes, collectées dans de multiples pays et sur de nombreuses années.

Nous aimerions donc mentionner les points ci-dessous, qui pourront peut être influencé ou ajouter des difficultés à notre analyse à venir :

* Rapport de l’année 2021: ce rapport fait suite au Covid-19 et à de nombreux confinements, qui ont fortement impacté le bien-être et la santé mentale de nombreuses populations.
* Échantillon de pays : ici les fichiers comptent 149 et 166. Ces chiffres sont inférieurs aux Etats (197) reconnus par l’Organisation des Nations unies en 2022).
* Échantillon de 1000 personnes par pays : ce chiffre est similaire pour chaque pays, mais ne couvre pas l’ensemble de la population nationale. Le World Happiness Report suggère d’avoir un échantillon de 2000 à 3000 personnes pour qu’il reflète réellement un pays.

# II – Présélection des variables

## 1 - Choix des colonnes à conserver ou supprimer

Nous avons vu auparavant que les colonnes ci-dessous représentaient les variables clefs et la variable cible. Nous avons décidé de les garder car elles sont pertinentes pour l’étude.

* Country name
* Life Ladder
* Log GDP per capita
* Social support
* Healthy life expectancy at birth
* Freedom to make life choices
* Generosity
* Perceptions of corruption

Nous avons également choisi de conserver les colonnes suivantes :

* **Year** : (Année)

Il s’agit de l’année de collecte des données.

La colonne contient donc des variables quantitatives (années).

Nous pensons également que cette colonne est essentielle, car elle nous permettra de faire des analyses d’évolution, ou année par année.

* **Positive effect** : (Aspect / Effet positif)

Il s’agit de la moyenne de trois mesures d’affect positif : bonheur, rire et plaisir. Ces mesures sont les réponses des trois questions suivantes, respectivement :

* Avez-vous ressenti un sentiment de bonheur pendant la journée d’hier ?
* Avez-vous ressenti un sentiment de plaisir pendant la journée d’hier ?
* Avez-vous ri hier ?

La colonne contient donc des variables quantitatives continues, qui représentent la moyenne des réponses sur une échelle de 0 à 10.

* **Negative effect** : (Aspect / Effet négatif)

Il s’agit de la moyenne de 3 mesures d’affect négatif : inquiétude, tristesse et colère. Ces mesures sont les réponses des trois questions suivantes, respectivement :

* Avez-vous ressenti un sentiment d’inquiétude pendant la journée d’hier ?
* Avez-vous ressenti un sentiment de colère pendant la journée d’hier ?
* Avez-vous ressenti un sentiment de tristesse pendant la journée d’hier ?

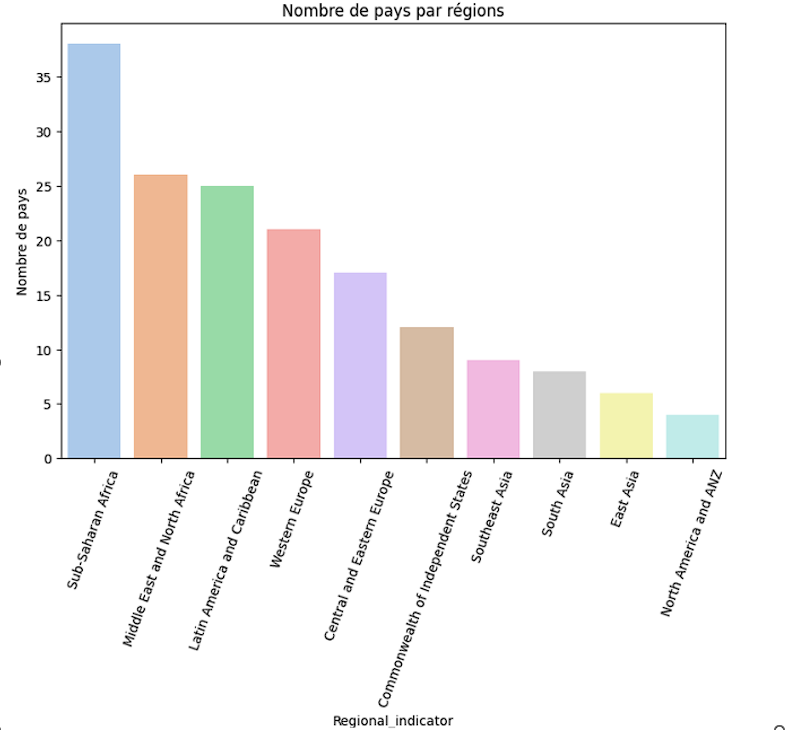
La colonne contient donc des variables quantitatives continues, qui représentent la moyenne des réponses sur une échelle de 0 à 10.

Nous ne sommes pas encore sûrs de l’importance des colonnes "Positive effect" et "Negative effect" et des variables associées.

Nous avons décidé de garder les colonnes suivantes du second fichier :

* Country Name
* Regional Indicateur : (Indicateur régional)

Les différents pays sont regroupés en 10 zones géographiques : Sub-Saharan Africa, Latin America and Caribbean, Western Europe, Central and Eastern Europe, Middle East and North Africa, Commonwealth of Independent States, Southeast Asia, South Asia, East Asia, North America and ANZ.



Nous remarquons que la répartition des pays est très inégale : une majorité des pays se situent dans la zone "Sub-Saharan Africa" (plus de 35) alors que la zone "North America and ANZ" en contient moins de 5.

Cette répartition inégale sera à prendre en compte dans nos futures analyses, car nous utiliserons uniquement le "Regional indicator" pour plus de visibilité, plutôt que tous les pays pour cette partie visualisation.

Enfin, nous avons décidé de supprimer les colonnes suivantes :

* upperwhisker
* lowerwhisker
* 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
* Standard error of ladder score
* Ladder score in Dystopia
* Dystopia + residual

Ces colonnes ne sont pas présentes dans le premier fichier et nous semblent peu pertinentes pour l’analyse. Nous ne sommes pas sûrs du mode de calcul ou de collecte de ces informations, et préférons les supprimer.

Pour conclure cette partie, nous créons un unique fichier, en concaténant les deux fichiers ensemble.

## 2 – Remplacement des valeurs manquantes

Le jeu contient différentes valeurs manquantes, notamment dans les colonnes 'Year', 'Regional Indicator', 'Positive affect' et 'Negative affect', suite à l’étape de concaténation.

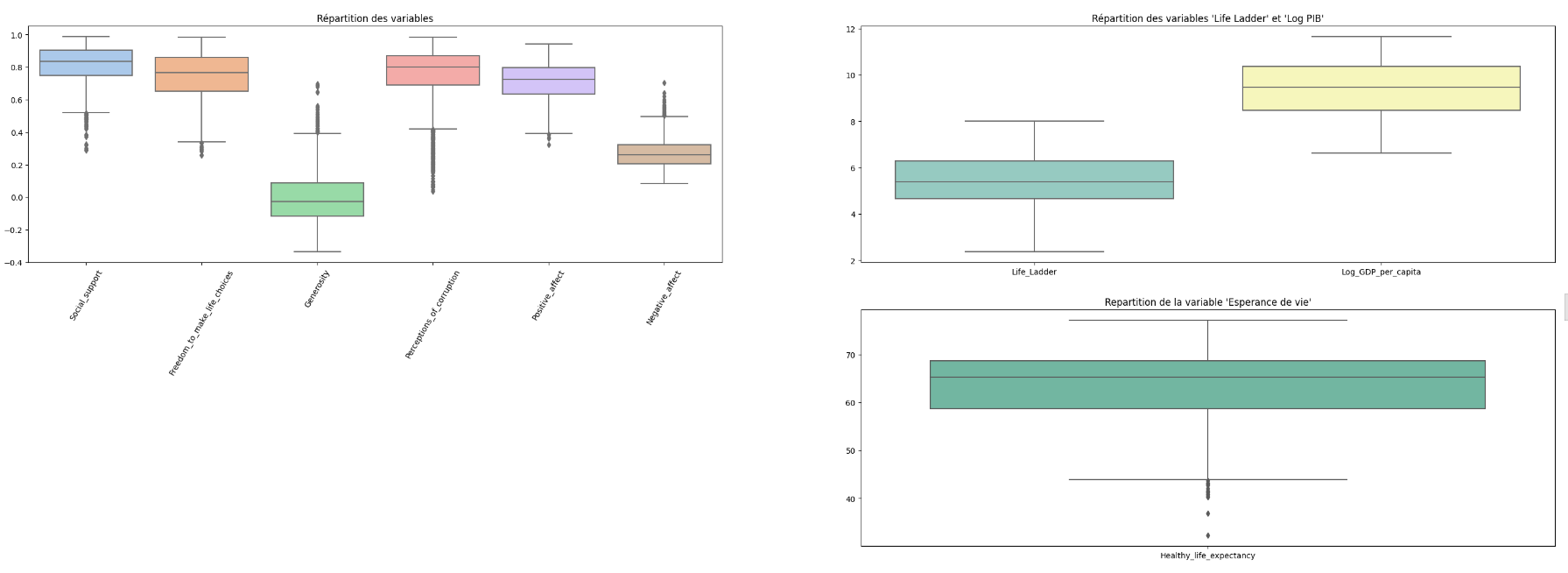
Pour la colonne 'Regional Indicator', cela était attendu, car elle n’était pas présente dans le jeu de données le plus complet, dans le fichier 'world-happiness-report.csv'. Il s’agit d’indicateurs régionaux pour regrouper des pays par zones. Nous devons donc relier les pays à leurs zones correspondantes.

Après avoir établi la liste des zones régionales, nous devons désormais relier les pays à leurs régions respectives, en fonction de celles qui étaient déjà présentes dans l’un des deux jeux de données originaux. Nous avons donc créer un dictionnaire (pays ; zone) en amont, afin de pouvoir l’utiliser pour remplacer les valeurs manquantes.

Enfin, il nous reste à remplacer les valeurs manquantes des colonnes 'Positive affect' et 'Negative affect'. Pour cela, nous avons décidé de les remplacer par la valeur médiane par pays sur les trois dernières années

# III - Visualisations et Statistiques

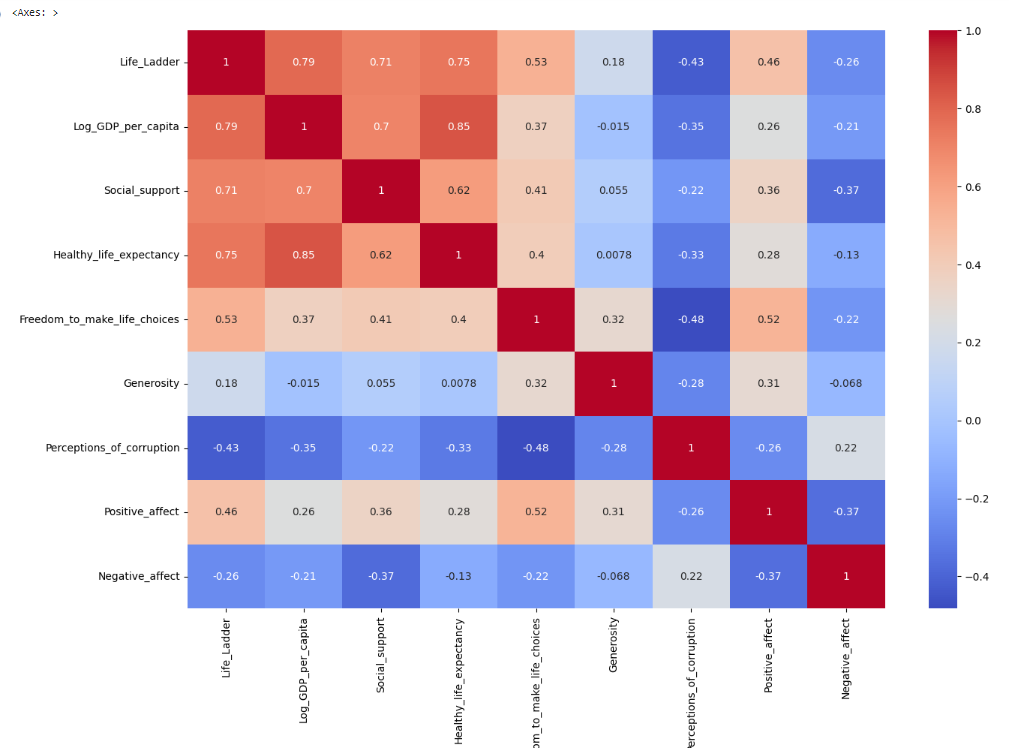
### Graphique 1 : Répartition des variables quantitatives



Il y a des extrêmes, nous devrons en tenir compte de ces extrêmes lors de l’étape de transformation des données, bien que ces variables soient le résultat de questions complètement différentes posées à l’échantillon, comme expliqué dans la première partie du projet.

Le graphique précédent nous a déjà orientés vers une étape de standardisation des données, mais ici, la technique recommandée semble plutôt être le Robust Scaling, qui est insensible aux valeurs extrêmes.

### Graphique 2 : Corrélation entres les variables quantitatives

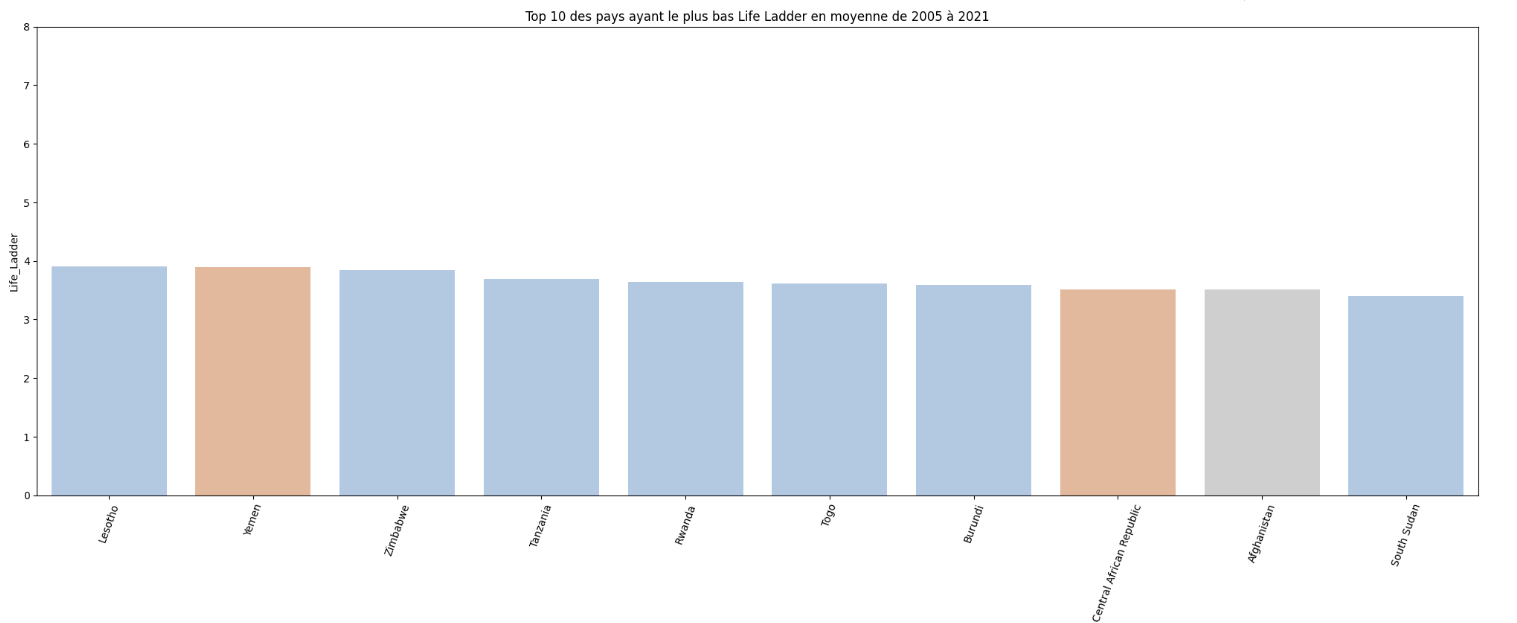
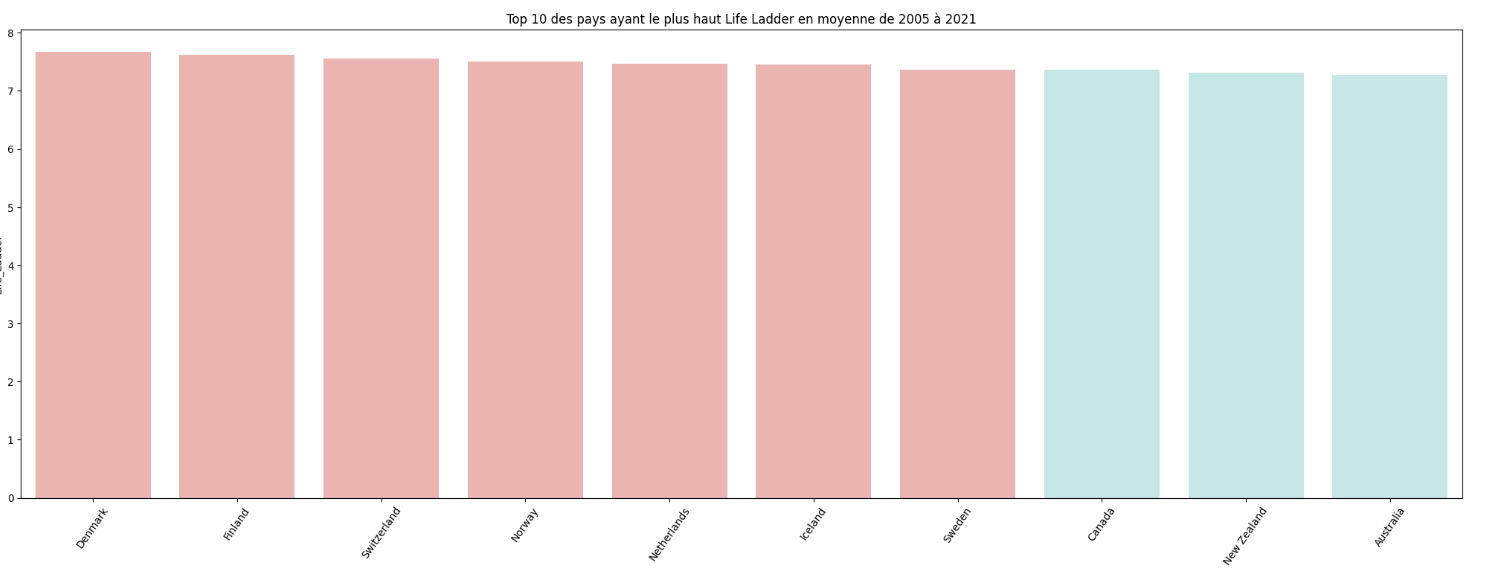
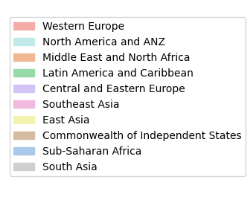


Il est d’usage en recherche de considérer qu’une valeur est forte lorsque sa valeur est supérieure à 0.7.

Les variables "Log GDP per capita" , "Social support" et "Healthy Life Expectancy" sont donc ici les plus positivement corrélées à la variable "Life Ladder", avec des scores de 0,79, 0,71 et 0,75.

Cela confirme leur impact positif sur la valeur cible.

### Graphique 3 : Top 10 des pays ayant la meilleure et la moins bonne échelle de vie

****

Il nous a paru essentiel de connaître les pays ayant la meilleure échelle de vie, et ceux ayant la plus basse.

Les pays ayant le plus haut score sont : Danemark, Finlande, Suisse, Norvège, Pays-Bas, Islande, Suède, Canada, Nouvelle-Zélande et Australie.

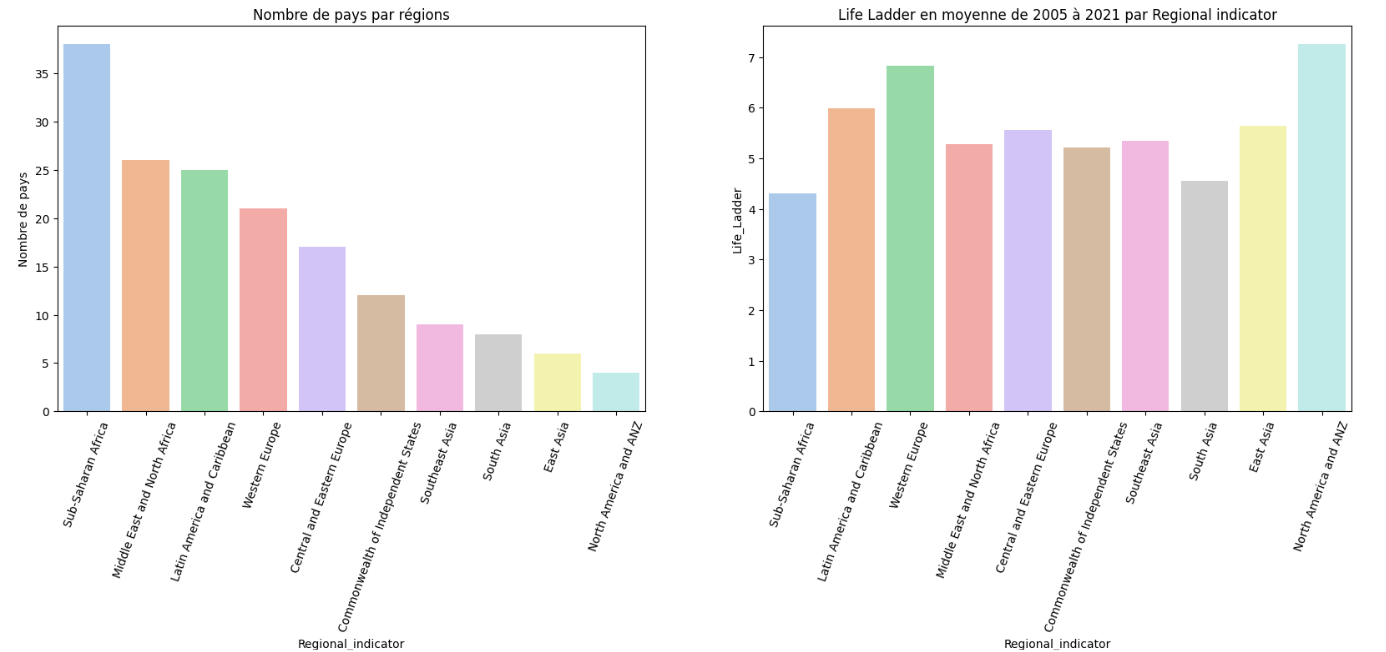
Les pays ayant le score le plus bas sont : Lesotho, Yemen, Zimbabwe, Tanzanie, Rwanda, Togo, Burundi, République Centrafricaine, Afghanistan, et Soudan du Sud.

Nous remarquons que deux régions se partagent le haut du classement : l’Europe de l’Ouest et l’Amérique du Nord avec l’Australie Nouvelle Zélande.

Deux régions occupent le bas du classement : l’Afrique Subsaharienne et le Milieu-Est et le Nord de l’Afrique.

Dans le Top 10 les valeurs sont comprises entre 7 et 8. Pour les 10 derniers pays les valeurs sont comprises entre 3 et 4.

### Graphique 4 : Moyenne du Life ladder pour la période de 2005 à 2021



Notre valeur cible étant la mesure du bonheur 'Life Ladder', nous trouvons pertinent de comparer cette répartition à cette valeur. Notre analyse s’étend donc de 2005 à 2021, en utilisant les moyennes annuelles de chaque pays.

"Sub-Saharan Africa" est la zone qui a le score le moins élevé, avec un 'Life Ladder' moyen à peine supérieur à 4. Nous aurions pu penser qu’avec un nombre important de pays, la zone "Sub-Saharan Africa" aurait le life ladder le plus important.

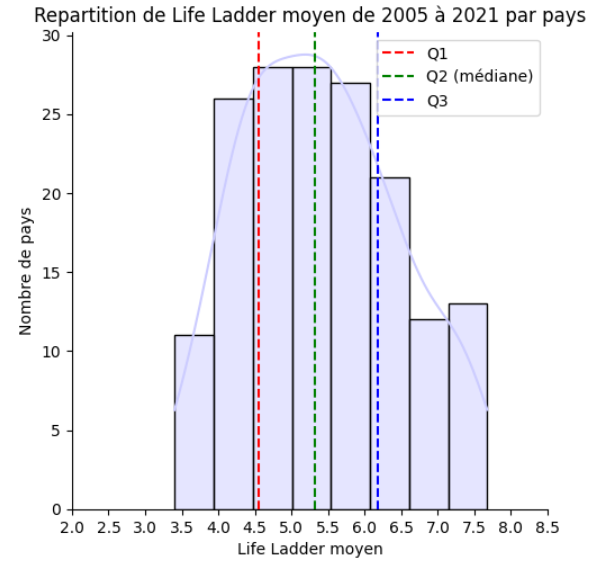
La zone "North America and ANZ" a le plus nombre de pays mais il a un ‘Life Ladder’ moyen supérieur à 7.

Le nombre de pays n’influence donc pas la variable "Life ladder". Cependant, les deux variables "Life Ladder' et "Regional Indicator" restent corrélées, ce que nous allons vérifier à l’aide du test statistique ci-dessous.

En effet nous disposons ici d’une variable qualitative 'Regional indicator' et d’une quantitative 'Life ladder'.

L’ANOVA nous renvoie un résultat p\_value **inférieur à 5%**. Nous pouvons donc affirmer que les modalités testées ont un effet statistique significatif sur la variable cible. La variable 'Life\_Ladder' est donc bien influencée par la variable 'Regional Indicator'.

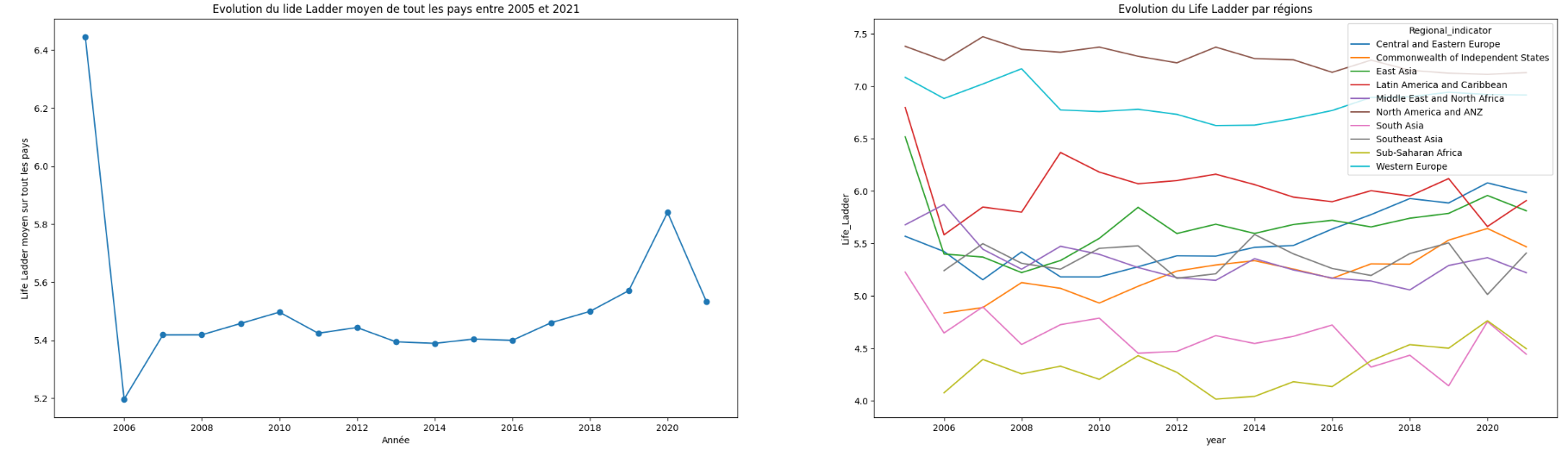
### Graphique 5 : Répartition de l’échelle de vie moyenne par pays et par quartiles



Avec l’aide des quartiles, nous pouvons constater que la majorité des pays ont un "Life Ladder" moyen compris entre 4.5 et 6.1 sur cette période, ce qui semble correct sur une échelle allant de 0 à 10. Les données sont regroupées, et il n’y a aucune donnée inférieure à 3 ou supérieure à 8.

La distribution des pays suit également une loi normale (courbe de Gauss). Cela est un indicateur pour nous orienter pour la partie transformation de données : nous pourrons utiliser la technique de standardisation pour une mise à l’échelle.

### Graphique 6 : Évolution de l’échelle de vie



Tout d’abord, nous pouvons noter différentes baisses :

* Une significative entre 2005 et 2006 sur les zones "Latin America Caribbean" et "East Asia"
* Une autre globale de 2005 à 2006 : le "Life ladder" moyen a également perdu environ 0.8 point.
* Une baisse en 2020, probablement due à la pandémie Covid-19

Suite à quelques recherches sur Internet, nous avons pu émettre les hypothèses suivantes :

* Des pays de l’Asie de l’Est ont été touchés par un tsunami de 2004. Ces pays ont donc souffert d’une hausse du stress, d’une augmentation de la pauvreté et d’une instabilité économique.
* Entre 2005 et 2006, le Venezuela a également été marqué par plusieurs événements politiques et économiques.
* Haïti a connu un coup d’État en 2004 qui a provoqué une instabilité politique pendant plusieurs années.

Il est intéressant de noter que ces événements ponctuels ont eu un impact significatif sur la qualité de vie des populations des régions concernées. Cela souligne l’importance de la stabilité politique et économique dans le maintien d’un niveau de vie satisfaisant pour les citoyens. Cette analyse permet de mieux comprendre l’influence de la richesse économique sur la qualité de vie des individus.

### Conclusion de l’ensemble des graphiques

Les graphiques ci-dessus confirment l’influence de différentes variables sur la variable cible "Life ladder" entre 2004 et 2021, notamment les variables "Log GDP per capita", "Social support" et "Healthy Life Expectancy".

Les corrélations nous ont également aiguillés vers des méthodes de standardisation comme le "Robust Scaling", qui est insensible aux valeurs extrêmes (présentes dans de nombreuses variables de notre jeu de données).

L’analyse étant terminée, la prochaine étape consiste donc à finaliser le jeu de données, avant de passer à la modélisation.

# IV – Finalisation du jeu de données

Nous allons diviser les données en deux ensembles: un ensemble X de variables prédictives et un ensemble y, la variable cible.

Ensuite, nous allons séparer le jeu de données en un jeu d’entraînement, et un jeu de test, de sorte que le jeu de test contienne 20% des données.​​

Nous devons ensuite remplir les valeurs manquantes du jeu d’entraînement et du jeu de test pour les variables numériques et catégorielles.

Les variables catégorielles ne sont pas exploitables pour la modélisation en machine learning, il est impératif de les transformer en valeurs numériques. Nous avons utilisé la méthode OneHotEncoder.

Nous avons dû rajouter un paramètre qui est "handle\_unknown=ignore", car certains pays du jeu de test ne font pas partie du jeu d’entraînement et cela peut générer des erreurs dans l’exécution du code.

Pour les variables numériques, nous avons choisi d’appliquer la méthode Simple Imputer qui permet d’imputer les valeurs manquantes au sein des données.

Pour simplifier le code, nous avons aussi appliqué un pipeline pour remplacer puis transformer ces variables numériques. Pour la standardisation de nos données, nous avons utilisé la méthode RobustScaler, insensible aux valeurs extrêmes.

# V - Modélisation

## 1 - Contexte et début de la modélisation

### 1 - a) Rappel

Pour rappel, ce projet consiste à effectuer une analyse approfondie des données collectées par le World Happiness Report. L’objectif pour cette seconde partie est donc de créer un modèle pour prédire les données pour l’année 2022.

Pour cette partie modélisation, il existe une large variété d’algorithmes de Machine Learning. Ici, puisque nous disposons d’une variable cible 'Life Ladder', nous sommes donc dans un **modèle d’apprentissage supervisé.**

Ici, notre variable cible 'Life Ladder' est le résultat d’un score, que nous pouvons considérer comme une **variable continue**. Par conséquent, nous pourrons utiliser des **algorithmes de régression.**

### 1 - b) Premiers essais et test d’hyperparamètres

Après plusieurs entraînements, nous nous sommes rendu compte qu’il était nécessaire de faire évoluer notre jeu de données précédent par l’ajout d’un set de validation. Il est courant d’évaluer les performances d’un modèle sur un jeu de données qu’il n’a pas encore vu.

Cela nous permettra de faire notre sélection de modèles avec de bonnes performances et d’identifier s’il y a un phénomène de surapprentissage.

Le jeu de validation nous permettra également d’obtenir les meilleurs hyperparamètres pour le futur modèle de prédiction.

Ensuite, nous avons donc décidé d’utiliser d’autres estimateurs. Nous avons commencé par tester un modèle RandomForest avec différents hyperparamètres grâce à ce pipeline. L’objectif était d’obtenir des résultats rapides, avec des temps de calcul limités.

Nous avons donc testé différents scénarios, avec des profondeurs de paramètres différents.

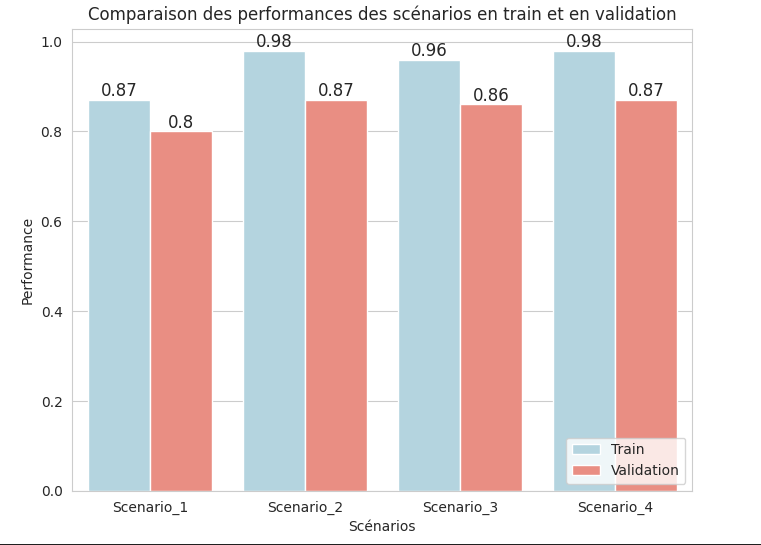
1-max\_depth=5, min\_samples\_split=3, n\_estimators=50

2-paramètre par défaut

3-max\_depth=10, n\_estimators=200

4-n\_estimators=200

Les temps d’exécutions varient : 12 secondes, 97 secondes, 105 secondes, 10 minutes, et environ 29 minutes.



Les temps d’exécutions varient : 12 secondes, 97 secondes, 105 secondes, 10 minutes, et environ 29 minutes.

Mis à part le premier essai, nous constatons ici qu’il n’y a pas de grande différence entre les scores, malgré les différents paramètres et temps d’exécution.

Cependant, en complexifiant les paramètres et en ajoutant de la profondeur, nous pouvons constater l’apparition de surapprentissage. C’est pourquoi nous avons décidé de ne pas continuer avec d’autres paramètres.

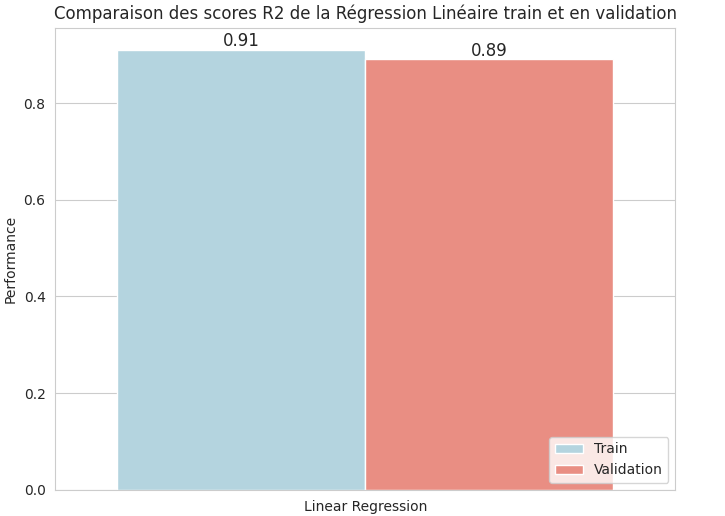
## 2 - Entraînement et validation des modèles

### 2 - a) Modèle LinearRegression

Ensuite, nous allons appliquer le modèle de régression linéaire.

Le Linear Régression est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

On obtient les résultats suivants :

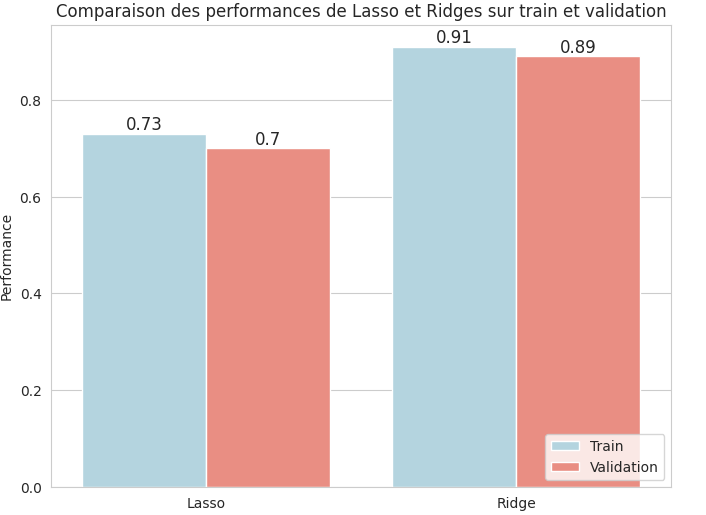


Ici, le modèle Linear Regression dispose d’excellents résultats : le score R2 est élevé à 0,91 et il y a peu d’écart entre le jeu d’entraînement et de validation. A noter également que pour ce modèle, il n’y a pas de phénomène de surapprentissage.

### 2 - b) Modèles LassoRegression et RidgeRegression

Plusieurs modèles de régression avec régularisation sont également disponibles. Nous allons nous concentrer ici uniquement sur les modèles Ridge et Lasso.

On obtient les résultats suivants :



Ici, c’est le modèle Ridge Regression qui dispose du meilleur score : le R2 est le plus élevé à **0,91** et il y a peu d’écart entre le jeu d’entraînement et de validation.

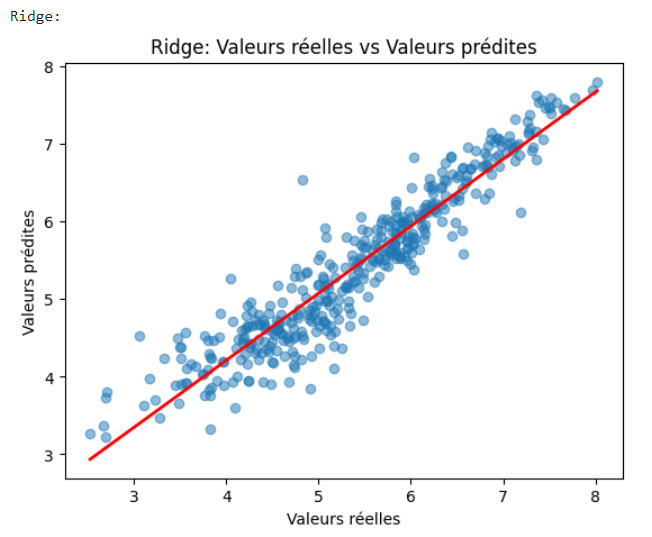
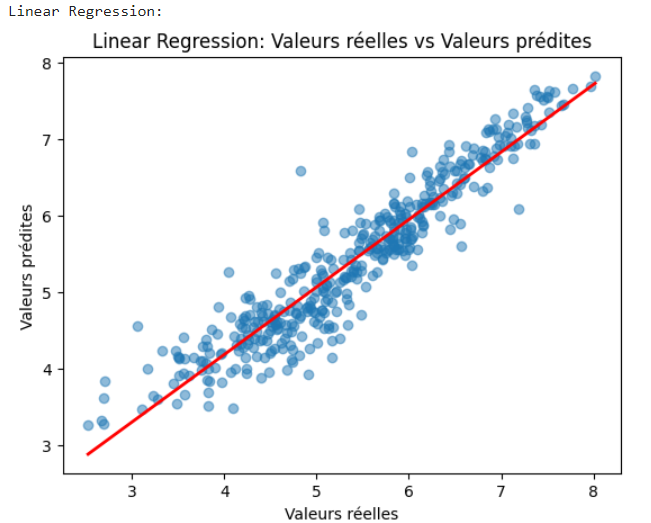
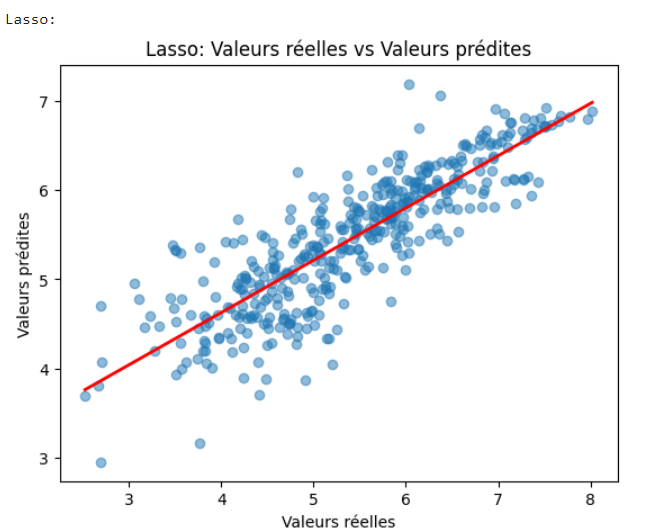
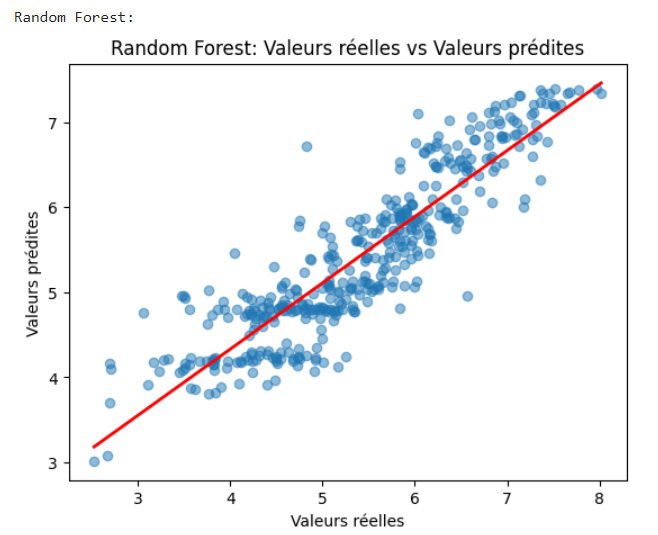
Le modèle **Ridge Regression donne exactement les mêmes résultats que ceux du modèle Linear Regression (0,91 et 0,89).**

Ils sont donc à privilégier face aux modèles Lasso Regression et Random Forest qui disposent de taux de R2 plus bas, laissant penser à une moins bonne prédiction.

## 3 - Fin des itérations de tests et sélection du modèle

### 3 - a) Analyse des valeurs prédites et des valeurs réelles

En plus du score R2, la répartition des valeurs prédites selon les valeurs réelles pour les différents modèles entrainés est également un indicateur pertinent dans la sélection d’un modèle de prédiction.



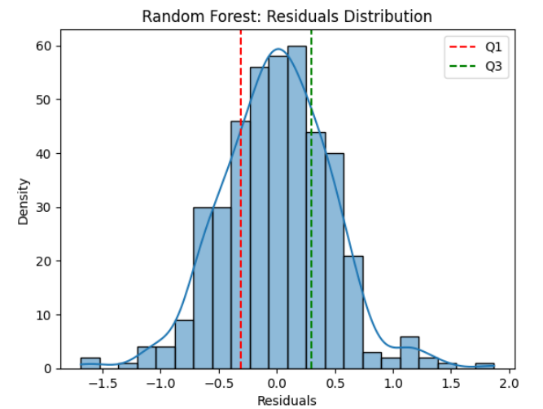
Sur chaque graphique, les points sont répartis le long de la courbe rouge de régression linéaire.

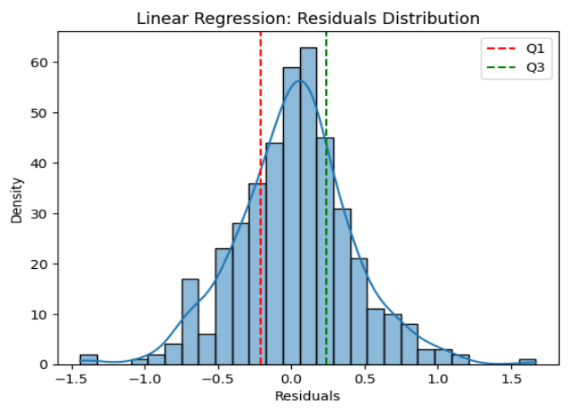
Cela confirme que notre approche est bonne et que les modèles de régression linéaire sont appropriés pour notre problématique.

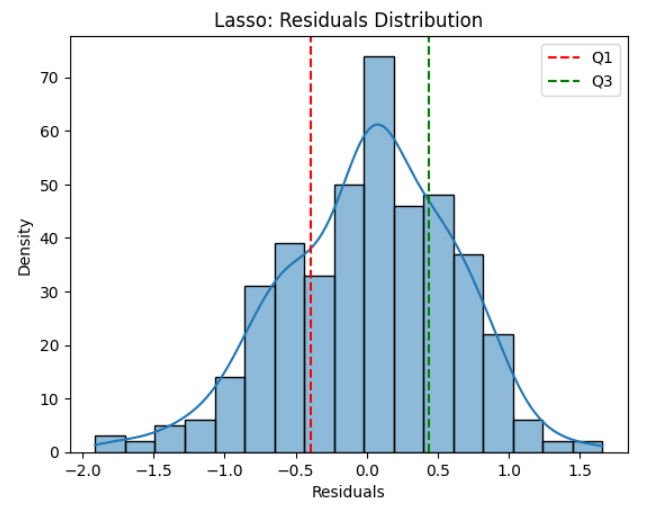
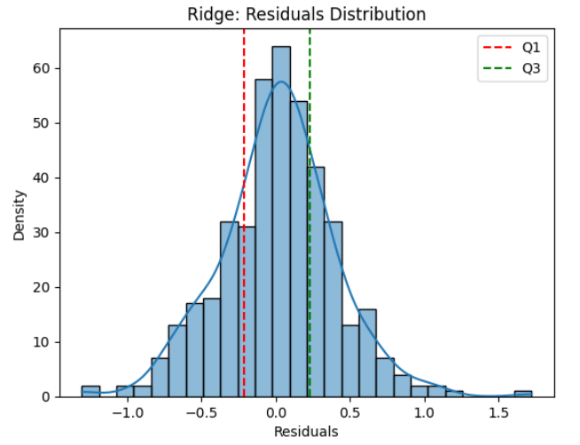
### 3 - b) Distribution des erreurs résiduelles

Enfin, nous souhaitons aussi étudier les erreurs résiduelles de chaque modèle, sous forme de graphiques.

En fonction de la distribution des erreurs, ceci peut également nous aiguiller sur la sélection d’un modèle de prédiction.





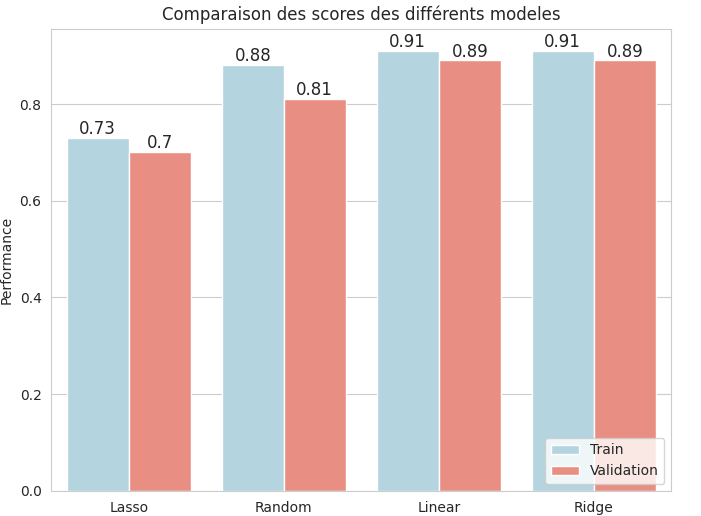


Nous remarquons que pour la Régression Linéaire et le Ridge Regression, la distribution des erreurs résiduelles suit la loi normale. Cela signifie que la majorité des erreurs est faible et que les erreurs importantes sont peu fréquentes.

D’autre part, nous pouvons observer que la distribution des erreurs est centrée autour du 0 pour ces deux modèles. La moyenne des erreurs est donc proche de 0, écartant la possibilité qu’il y ait un biais systématique dans l’usage des modèles Linear Régression et Ridge Régression.

### 3 - c) Récapitulatif et conclusion

Pour plus de visibilité, nous souhaitons également comparer visuellement les modèles et les scores obtenus entre eux.



Ce graphique nous permet de clairement visualiser le résultat de l’ensemble des modèles.

Ici, comme précédemment expliqués, deux modèles affichent les scores les plus élevés: le Linear Régression et le Ridge Régression. Il nous faut donc désormais choisir entre ces deux modèles.

Au vu des analyses précédentes, et notamment des graphiques, nous avons décidé de choisir le modèle Linear Régression.

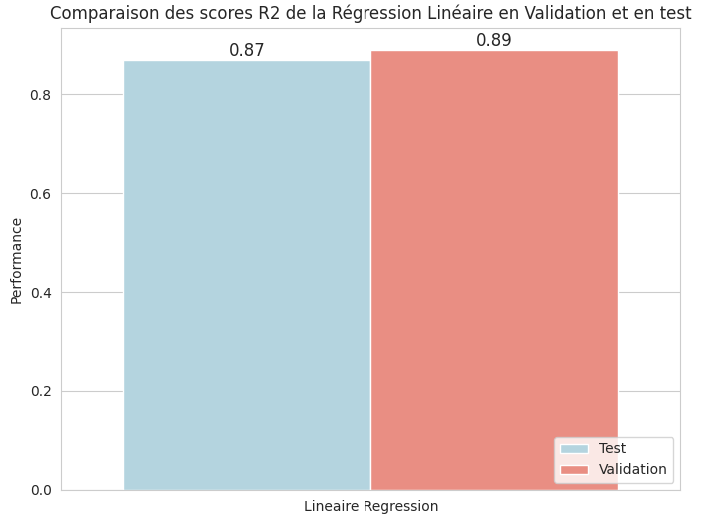
En effet, le modèle dispose d’un score R2 élevé et ne crée pas de phénomène de surapprentissage. Les deux précédents graphiques ont également confirmé la cohérence du Linear Régression.

Enfin, le Linear Régression est également très approprié pour les jeux de données simples, ce qui est le cas pour notre jeu de données actuel.

## 4 - Évaluation du modèle sur le jeu de test

### 4 - a) Application sur le jeu de test

Au vu des précédentes informations, nous avons donc finalement privilégié le modèle Linear Régression.



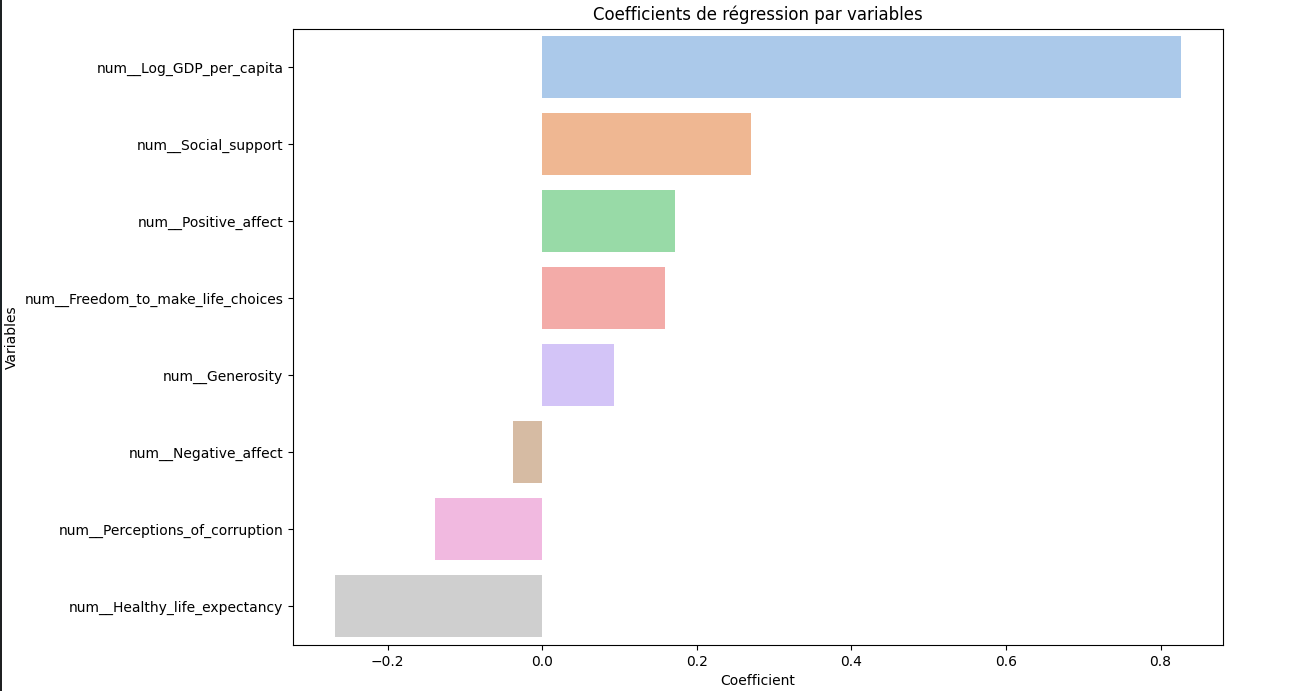
Nous disposions d’un score R2 de **0.89** sur le jeu de validation Linear Regression.

Le score R2 actuel pour le jeu de test est donc très proche avec **0.87**.

Cela nous laisse penser que ce modèle fonctionne parfaitement et pourrait permettre d’obtenir des prédictions précises du "Life ladder" pour les prochaines années.

### 4 - b) Coefficients et identification des variables clefs

Pour mieux comprendre et interpréter le résultat précédent du modèle de régression linéaire, nous pensions qu’il était nécessaire de s’intéresser aux coefficients et de déterminer les variables qui influencent ce modèle.



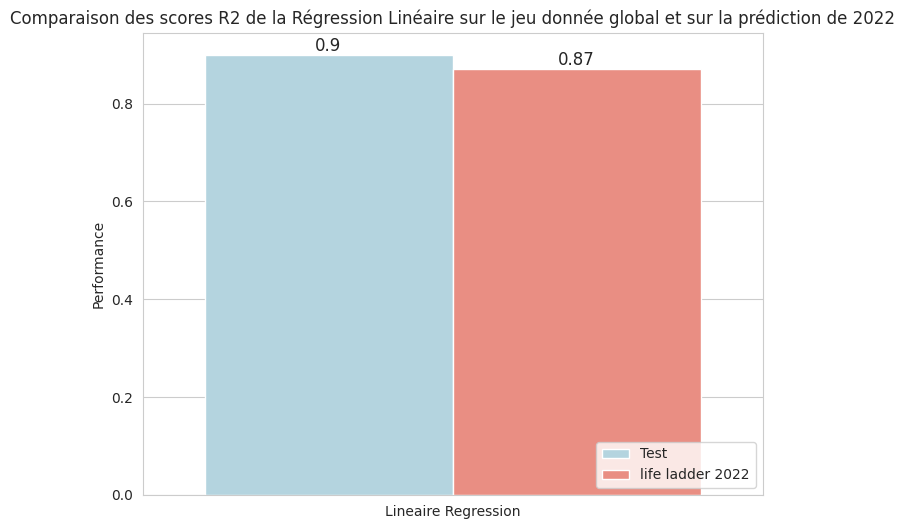
Ici, le graphique indique que les variables "Log GDP per capita" et "Social Support" ont une influence positive importante sur la variable cible "Life ladder". Cette influence avait déjà été mentionnée dans la première partie.

Nous pouvons également noter que certaines variables comme "Healthy life Expectancy" influencent négativement la variable cible "Life ladder", mais à moindre mesure que les variables positives précédente.

# VI – Application du modèle sur de nouvelles données

Au cours du projet, un nouveau rapport a été rendu disponible. Il s’agit du rapport World Happiness Report 2023 (publié le 20 mars) avec l’ensemble des données de l’année 2022.

Nous allons donc effectuer une prédiction avec notre Linear modèle sur l’année 2022 et comparer les résultats obtenus à ceux présents dans le nouveau rapport 2023.



Nous obtenons alors un score R2 de **0,87** pour cette nouvelle prédiction, ce qui confirme que notre modèle fonctionne et qu’il est en accord avec les nouvelles données de 2022.

# VII - Conclusion

## 1 - Conclusion métier

Pour ce projet, nous avons effectué une analyse approfondie des données collectées par le World Happiness Report.

Nous avons donc cherché à estimer l'influence de ces différentes mesures socio-économiques autour de la santé, l’éducation, la corruption, l’économie, l’espérance de vie sur le bonheur.

En conclusion, ce projet d'étude nous a permis de mieux comprendre les facteurs qui influencent le bonheur des citoyens à travers le monde. Nous avons identifié le PIB par habitant comme étant le facteur clé qui explique le mieux la variation du classement du World Happiness. Les autres facteurs qui ont également une forte influence sur le Life Ladder sont le support social et l'espérance de vie en bonne santé.

Après avoir nettoyé notre base de données et effectué des visualisations, nous avons pu développer un modèle de prédiction à l’aide de technique de machine Learning.

Il offre des perspectives intéressantes pour la mise en place de politiques publiques et des recommandations pour les entreprises et les organisations internationale. Il permet d’aiguiller les associations vers les régions les plus nécessiteuses.

Cependant, il est important de noter que notre étude a ses limites, notamment en termes de taille d'échantillon. Bien que nous ayons utilisé un ensemble de données considérable, avec environ 1000 sondages par pays, il est recommandé d'utiliser un échantillon minimum de 2000 pour obtenir des résultats plus fiables.

## 2 - Retour d’expérience

Pour conclure, nous sommes tous les trois satisfaits de la progression du projet et du compte-rendu actuel. Nous avons apprécié collaborer sur ce sujet. Chacun d’entre nous a pu apporter son expertise et son avis tout au long du projet.

Nous avons dû être attentif à notre organisation afin de prendre en compte le décalage horaire de Jérémie situé à l’île de la Réunion d’une part, et notre rythme d’apprentissage différent d’autre part. Nous étudions les cours en différé. Nous devions faire en sorte de consacrer assez de temps au projet sans délaisser les modules et les certifications.

Pour cela, nous avons organisé différentes sessions de travail communes, plusieurs fois par semaine. Nous avions également des temps de travail asynchrone, en utilisant différents fichiers Google Colaboratory et un Google drive partagé. Cette organisation nous a permis de tenir l’organisation et les délais suggérés par Datascientest, semaine par semaine.

Lors de ce projet, nous avons eu la chance de ne pas rencontrer de difficulté majeure.

Nous avons rencontré quelques problèmes d’exécution de code lors des entraînements de modélisation.

Pour nos premières itérations de test, les profondeurs des paramètres étaient trop importantes. Les délais d’attentes étaient très longs. Nous les avions interprété à tort comme une erreur de code. Pour nos tests suivants, le temps d’exécution le plus long était de 1h 40 minutes.

## 3 - Axes d’amélioration

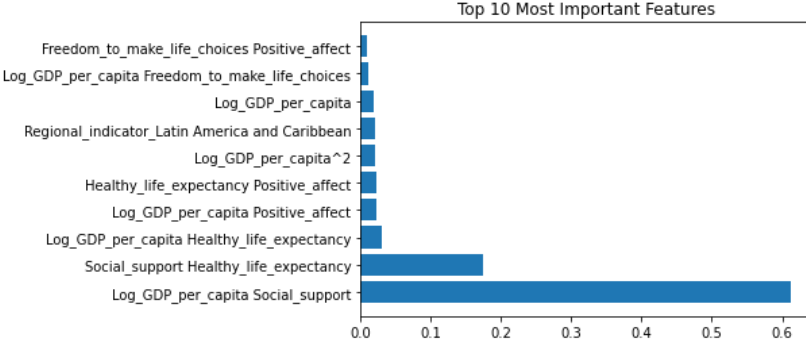
Nous sommes également conscients que des axes d’amélioration sont encore possibles.

* Nous aurions par exemple, pu appliquer d’autres modèles de régression avec

régularisation comme les modèles ElasticNet ou bien SVR.

* Dans nos premiers essais, nous avons également essayé d’appliquer une

combinaison de modèles polynomiales à des régressions linéaires, qui est une autre méthode permettant de complexifier le jeu de donnée. Cette approche a montré des regroupements de variables intéressantes qui pourraient encore être exploré et contribuer à affiner l’identification des facteurs clés, à savoir le PIB et le soutien social d’une part, et l’espérance de vie en bonne santé et le soutien social d’autre part.



* Puisqu’il s’agit de sujet d’actualité, couvert régulièrement par les médias, nous aurions

pu utiliser du WebScrapping pour étoffer notre analyse. Certains sujets abordés lors de la formation n’ont également pas été utilisés lors de notre projet.

* De nombreuse analyse de ce sujet sont déjà en ligne. Avec plus de temps, nous

aurions aimé pouvoir les étudier plus en détail et les comparer avec les nôtres.

* Suite aux différents tests, nous voulions également préciser que l’ensemble du code

de modélisation peut être groupé et simplifié par une boucle. Voir le code en annexe.

* Pour comparer les modèles et leurs scores, nous avons également testé la méthode

de validation croisée. Voir le code en annexe.

## 4 - Rappel des sources

<https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021>

<https://worldhappiness.report/>

<https://worldhappiness.report/faq/>

<https://happiness-report.s3.amazonaws.com/2021/Appendix1WHR2021C2.pdf>

<https://www.un.org/fr/about-us/member-states>

# VIII - Annexes

<https://github.com/NicolasCuv/World_happiness>