



Projet : Analyse du bien-être sur Terre

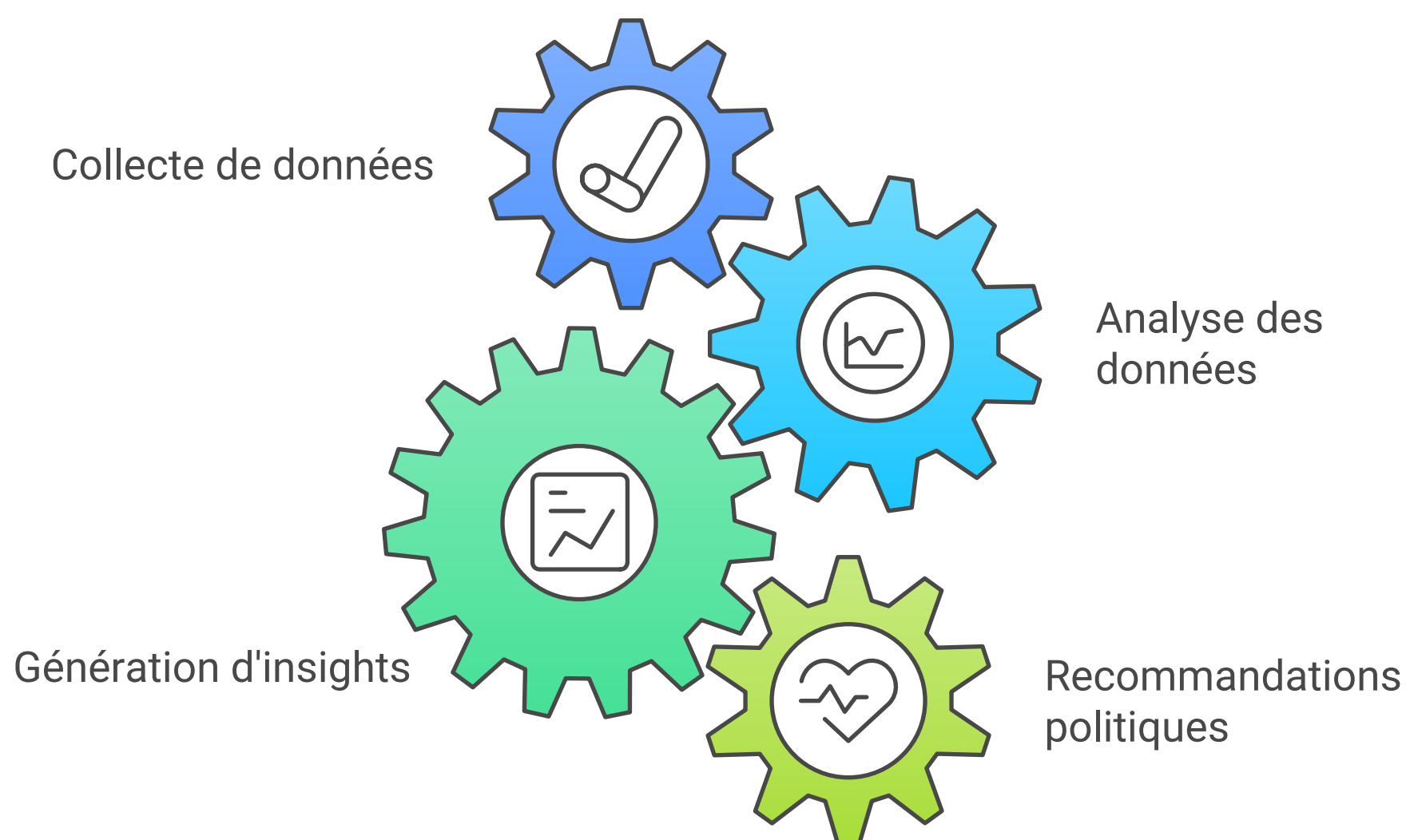


RAPPORT D'EXPLORATION



Description du Projet : **Analyse du bien-être mondial à partir des données du World Happiness Report**

Étapes de l'analyse du bien-être mondial



Ce projet vise à conduire une analyse approfondie des données du World Happiness Report afin d'évaluer le bonheur des pays du monde en utilisant une variété d'indicateurs socio-économiques tels que la santé, l'éducation, la corruption, l'économie et l'espérance de vie.

L'objectif principal est de présenter ces données à travers des visualisations interactives bien conçues tout en identifiant les combinaisons de facteurs qui expliquent pourquoi certains pays sont mieux classés que d'autres en termes de bonheur.

Présentation des Données :

Deux jeux de données ont été fournis au format csv.

1- « world-happiness-report.csv » : données du World Happiness Report de 2005 à 2020

2- « world-happiness-report-2021 » : données du World Happiness Report de 2021

Présentation des variables du jeu de données « world-happiness-report.csv » :

Country name : Nom du pays.

year : Année de l'enquête.

Life Ladder : Indicateur du bonheur de vie [plus la valeur est élevée, plus le bonheur est élevé]. Les évaluations constituent la base du classement annuel du bonheur. Ils sont basés sur les réponses à la question principale d'évaluation de la vie selon l'échelle de Cantril. L'échelle de Cantril demande aux personnes interrogées de penser à une échelle, la meilleure vie possible pour eux étant un 10 et la pire vie possible étant un 0. On leur demande ensuite d'évaluer leur propre vie actuelle sur cette échelle de 0 à 10. Les classements sont établis à partir d'échantillons représentatifs à l'échelle nationale (pour chaque pays 1 000 personnes par an) sur une période de trois ans.

Log GDP per capita : PIB par habitant enregistré sur une échelle logarithmique.

Social support : Mesure du soutien social perçu. Le soutien social (ou le fait d'avoir quelqu'un sur qui compter en cas de problème) est la moyenne nationale des réponses binaires (0 ou 1) à la question du PRG Si vous avez eu des problèmes, avez-vous des parents ou des amis sur lesquels vous pouvez compter pour vous aider chaque fois que vous en avez besoin, ou non

Healthy life expectancy at birth : Espérance de vie en bonne santé à la naissance. Données extraites du référentiel de données de l'Observatoire mondial de la santé de l'Organisation mondiale de la santé (OMS)

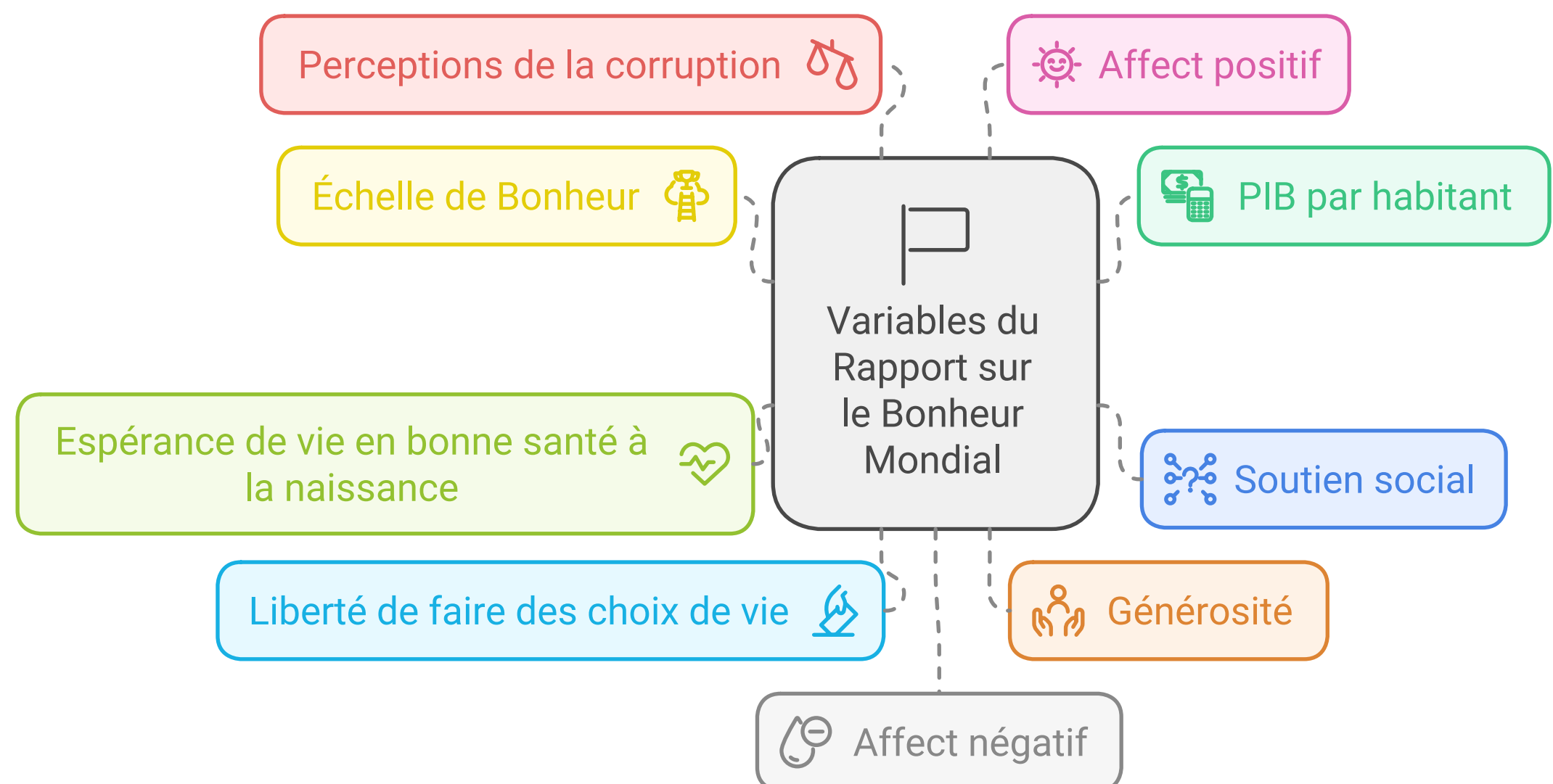
Freedom to make life choices : Mesure de la liberté de faire des choix dans la vie. La liberté de faire des choix de vie est la moyenne nationale des réponses à la question du PRG Êtes-vous satisfait ou insatisfait de votre liberté de choisir ce que vous faites de votre vie ?

Generosity : Mesure de la générosité de la population. La générosité est le résidu de la moyenne nationale régressive de réponse à la question du PRG Avez-vous donné de l'argent à un organisme de bienfaisance au cours du dernier mois ? sur le PIB par habitant.

Perceptions of corruption : Mesure de la perception de la corruption dans le pays. La mesure est la moyenne nationale de l'enquête en réponse à deux questions du GWP : la corruption est-elle répandue dans l'ensemble du gouvernement ou non et la corruption est-elle répandue dans les entreprises ou non ? La perception globale n'est que la moyenne des deux réponses 0 ou 1. Dans le cas où la perception de la corruption gouvernementale est absente, nous utilisons la perception de la corruption des entreprises comme perception globale. La perception de la corruption au niveau national n'est que la réponse moyenne de la perception globale au niveau individuel

Positive affect : Mesure de l'affect positif. L'affect positif est défini comme la moyenne de trois mesures d'affect positif dans le GWP : le bonheur, le rire et le plaisir dans les vagues 3 à 7 du Gallup World Poll. Ces mesures sont les réponses aux trois questions suivantes, respectivement : Avez-vous ressenti les sentiments suivants pendant une grande partie de la journée ? Qu'en est-il du bonheur ? Avez-vous beaucoup souri ou ri hier ? Avez-vous ressenti les sentiments suivants pendant une grande partie de la journée d'hier ? Qu'en est-il du plaisir ? Les vagues 3 à 7 couvrent les années 2008 à 2012 et un petit nombre de pays en 2013. Pour les vagues 1 et 2 et celles à partir de la vague 8, un affect positif est défini comme la moyenne du rire et du plaisir uniquement, en raison de la disponibilité limitée du bonheur

Negative affect : Mesure de l'affect négatif. L'affect négatif est défini comme la moyenne de trois mesures d'affect négatif dans le PRG. Il s'agit de l'inquiétude, de la tristesse et de la colère, respectivement les réponses à Avez-vous ressenti les sentiments suivants pendant une grande partie de la journée d'hier ? Qu'en est-il de l'inquiétude ? Avez-vous ressenti les sentiments suivants pendant une grande partie de la journée d'hier ? Qu'en est-il de la tristesse ? Avez-vous ressenti les sentiments suivants pendant une grande partie de la journée d'hier ? Qu'en est-il de la colère ?



I. Objectifs et Résultats Attendus

Notre analyse vise à explorer les données du World Happiness Report pour évaluer le bonheur des pays du monde en fonction de différents indicateurs socio-économiques inclus dans le jeu de données de base. Nous souhaitons l'enrichir avec de nouvelles variables telles que les données de changement climatique. Afin de savoir si elles peuvent influencer sur Life Ladder.

Nous prévoyons également d'ajouter une variable « indicateur de la région » afin de pouvoir faire des groupements et une analyse par région.

Nous avons également l'intention de construire un modèle prédictif pour estimer le score de bonheur en utilisant différentes techniques d'apprentissage automatique telles que la régression ou la classification.

II. Les différentes étapes du projet

1) La compréhension des données

2) La collecte de nouvelles données relatives au changement climatique « climate_change_indicators.csv »

Source : [Climate change Indicators \[kaggle.com\]](https://www.kaggle.com/datasets/ClimateChangeIndicators). Obtenu à partir de diverses sources en ligne liées aux indicateurs des changements climatiques tels que worldbank.org et climatedata.imf.org. Cet ensemble de données contient des indicateurs de changement climatique pour différents pays avec leurs codes associés [ISO2 ET ISO3]. La mesure a été mise à jour chaque année jusqu'en 2022 à partir de 1961.

72 variables

ObjectId : ID

Country : Nom du pays

ISO2 : Codes pays de deux lettres définis dans la norme ISO 3166 [partie ISO 3166-1] publiée par l'Organisation internationale de normalisation [ISO] pour représenter les pays

ISO3 : ISO 3166-1 est une norme internationale de codification des pays - codes sur trois lettres, permettant une association visuelle avec le nom usuel du pays

Indicator : Variation de température par rapport à une climatologie de référence, correspondant à la période 1951-1980

Unit : Unité : Degree Celsius

Source : Food and Agriculture Organization of the United Nations (FAO). 2022. FAOSTAT Climate Change, Climate Indicators, Temperature change. License: CC BY-NC-SA 3.0 IGO. Extracted from: <https://www.fao.org/faostat/en/#data/ET>. Accessed on 2023-03-28.

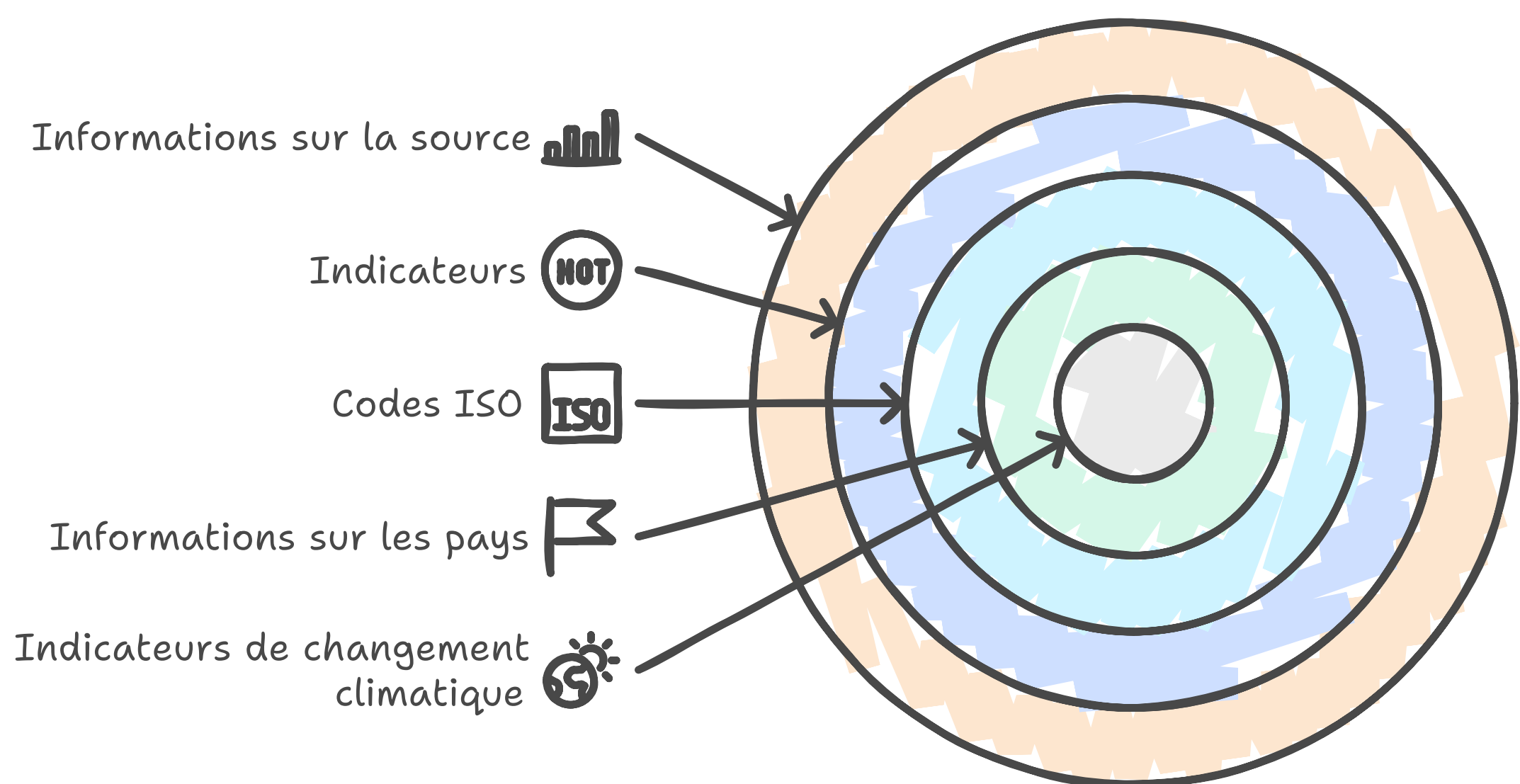
CTS_Code : ECCS

CTS_Name : Surface Temperature Change

CTS_Full_Descriptor : Environment, Climate Change, Climate Indicators, Surface Temperature Change

62 variables de F1962 à F2022 : mesures prises de 1962 à 2022

Structure de l'ensemble de données sur les indicateurs de changement climatique



Nous utiliserons les variables Country et ISO3 qui nous serviront de variables de correspondance pour merger avec le dataset world-happiness-report.csv. Ainsi que les variables de F2005 à F2022.

Ajout d'une variable « Région », chaque pays sera affecté à une région selon sa position géographique. Cela permettra d'étudier si le score du bonheur a un lien avec la région géographique.

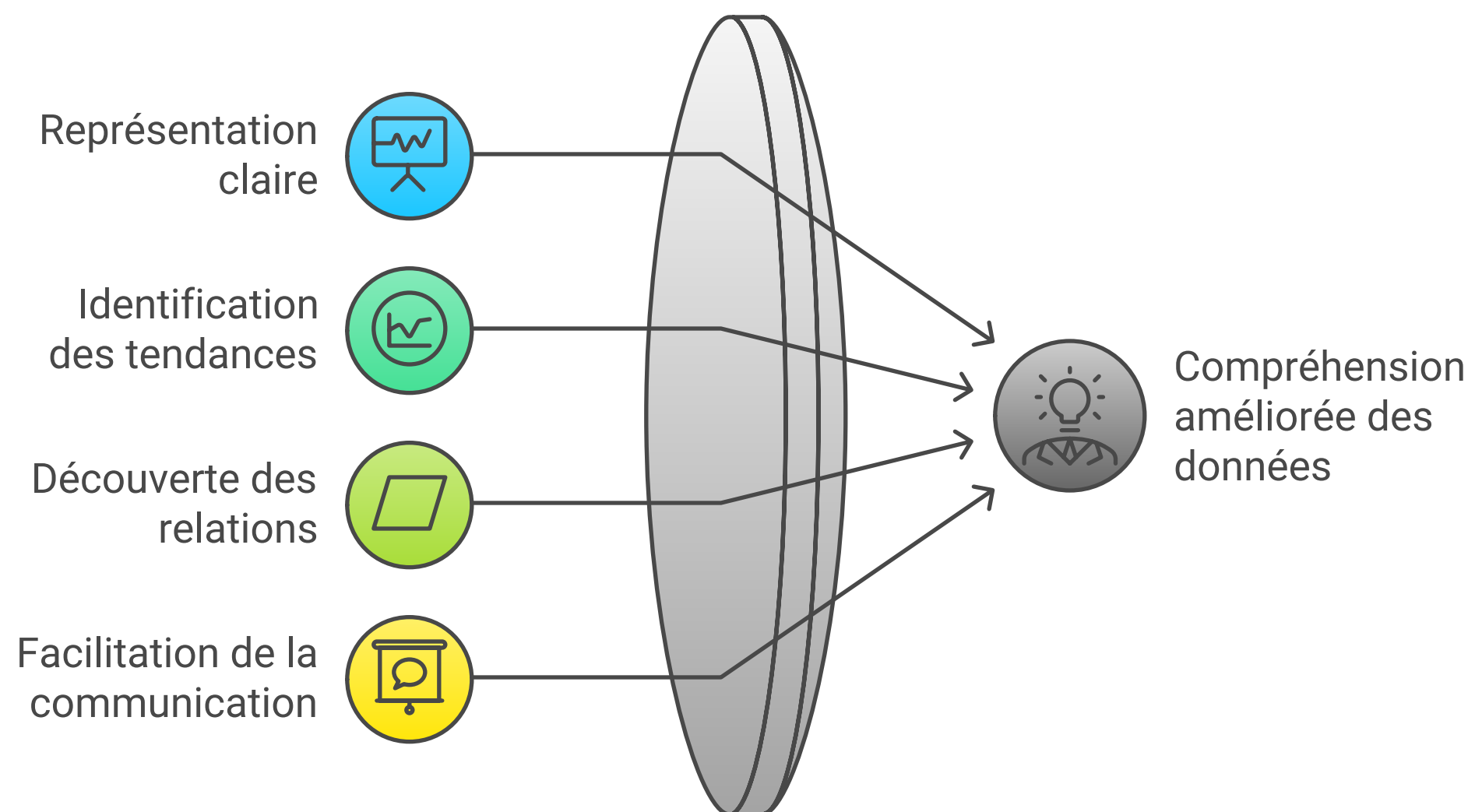
3) Intégration des nouvelles données

Concaténer les fichiers « world-happiness-report.csv » + « climate_change_indicators.csv » afin d'avoir le « Fichier_final.csv ».

Le dataset qui sera notre jeu de données pour la suite du projet sera le « Fichier_final.csv ».

III. Les visualisations

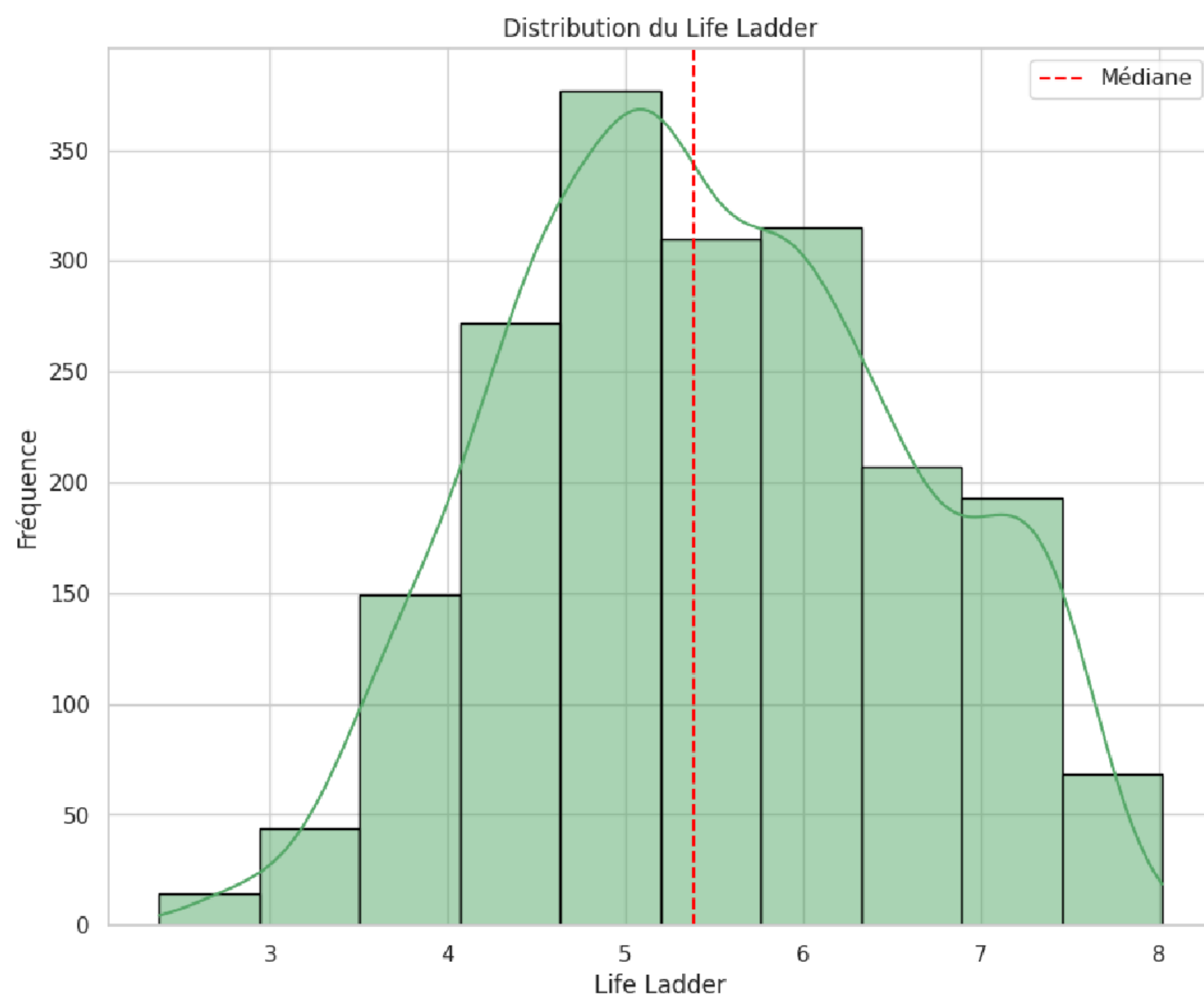
Visualisations pour la compréhension des données



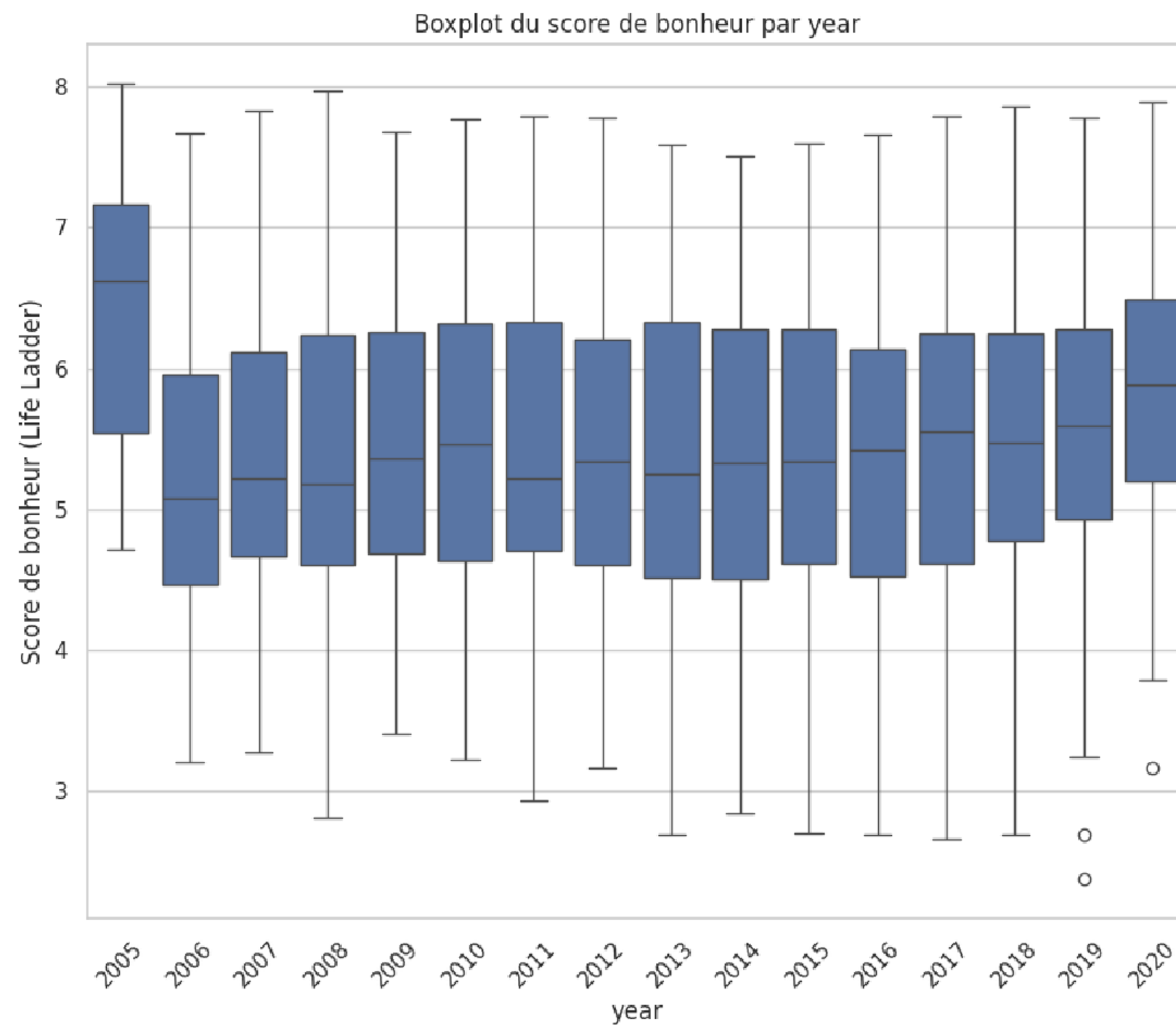
A) Approche générale

1) Visualisation de la distribution du score du bonheur

La fréquence des différents scores de bonheur sur une échelle de 0 à 8, où 0 représente le moins heureux et 8 le plus heureux. La forme du graphe est asymétrique, avec une queue plus longue vers la gauche. Cela indique qu'il y a plus de pays qui ont un score de bonheur inférieur à la moyenne que supérieur à la moyenne.



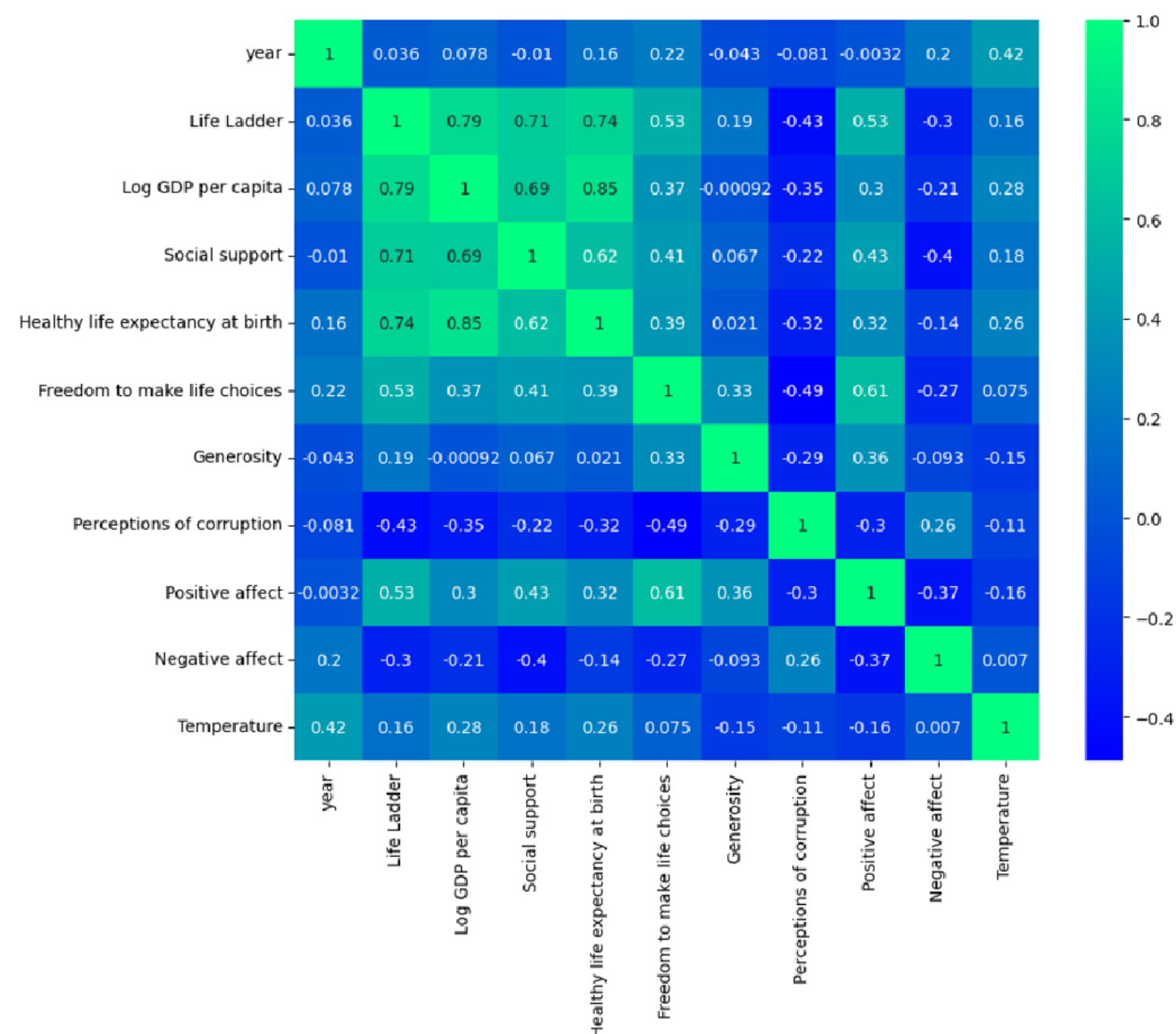
2) Evolution du score du bonheur par année



Constats principaux :

- Les scores de bonheur semblent avoir augmenté au fil des années, avec une légère variation d'une année à l'autre.
- La médiane (ligne au milieu de chaque boîte) semble également augmenter progressivement.
- Les valeurs extrêmes en 2019 et 2020 pourraient être dues à des circonstances exceptionnelles (telle la pandémie de COVID-19).

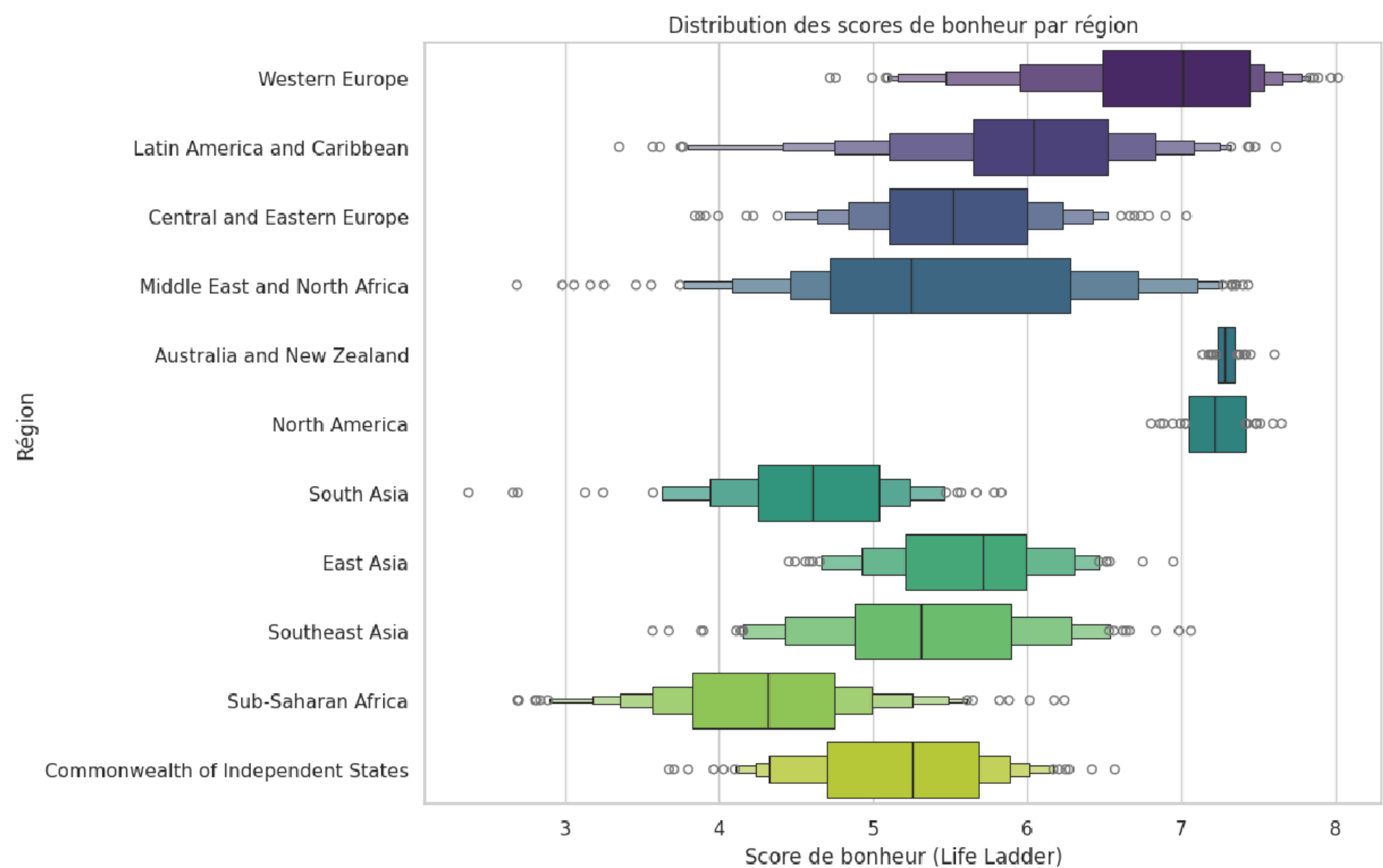
3) Matrice de corrélation des variables



Constats principaux :

- Les variables ayant les corrélations positives les plus élevées avec la variable “Life Ladder” sont les suivantes : “Log GDP per capita” [coefficient de 79%], “Healthy Life Expectancy at birth” [coefficient de 74%] et “Social support” [coefficient de 71%] et “Freedom to make life choices” [coefficient de 53%]. Dans la suite de notre analyse, nous étudierons de façon plus détaillée les relations existantes entre ces différentes variables et notre variable cible “Life Ladder”. Cela implique, dans le cas du produit intérieur brut par habitant [“Log GDP per capita”] par exemple, que plus il est élevé et plus le score de bonheur est élevé, et inversement.
- Le coefficient entre « Life Ladder » et « Temperature » est de 18%, il y a une faible corrélation entre ces deux variables. Cela peut indiquer que le réchauffement climatique n’a pas de grande influence sur le score du bonheur.
- Le coefficient entre “Freedom to make life choices” et “Perceptions of corruption” est de -44%, il y a une corrélation négative modérée entre ces deux variables. Cela implique que plus la liberté de choix est élevée, plus la perception de la corruption est faible, et inversement.
- Le coefficient entre “Generosity” et “Log GDP per capita” est de 0,0092% ce qui signifie qu’il n’y a quasiment pas de corrélation entre ces deux variables. Cela implique que le niveau de générosité n’est pas lié au niveau de richesse, et qu’il peut varier indépendamment.

4) Analyse de la distribution par Région à l’aide des Boxplot



Constats principaux :

- La région « Australie et la Nouvelle-Zélande possède le score de bonheur moyen le plus élevé, avec environ 7.2. Sa boîte à moustaches est étroite et symétrique. Les scores de bonheur sont peu dispersés et proches de la moyenne.
- Il n'y a pas de valeurs aberrantes, ce qui signifie que tous les pays de cette région ont un niveau de bonheur similaire.
- La région « Afrique subsaharienne » a le score de bonheur moyen le plus bas, avec environ 4,5. Sa boîte à moustaches est large et asymétrique. Les scores de bonheur sont très dispersés et plus faibles que la moyenne. Il y a plusieurs valeurs extrêmes, ce qui signifie que certains pays de cette région ont un niveau de bonheur très différent des autres.
- La région « Europe occidentale » a un score de bonheur moyen élevé, avec environ 5,1. Sa boîte à moustaches est étroite et légèrement asymétrique.

Les scores de bonheur sont peu dispersés et légèrement supérieurs à la moyenne. Il y a quelques valeurs extrêmes, ce qui signifie que certains pays de cette région ont un niveau de bonheur plus bas ou plus haut que les autres.

B) Approche détaillée

Afin d'approfondir notre analyse, nous proposons, sur la base de matrice de corrélation présentée précédemment, de réaliser 4 analyses bivariées entre :

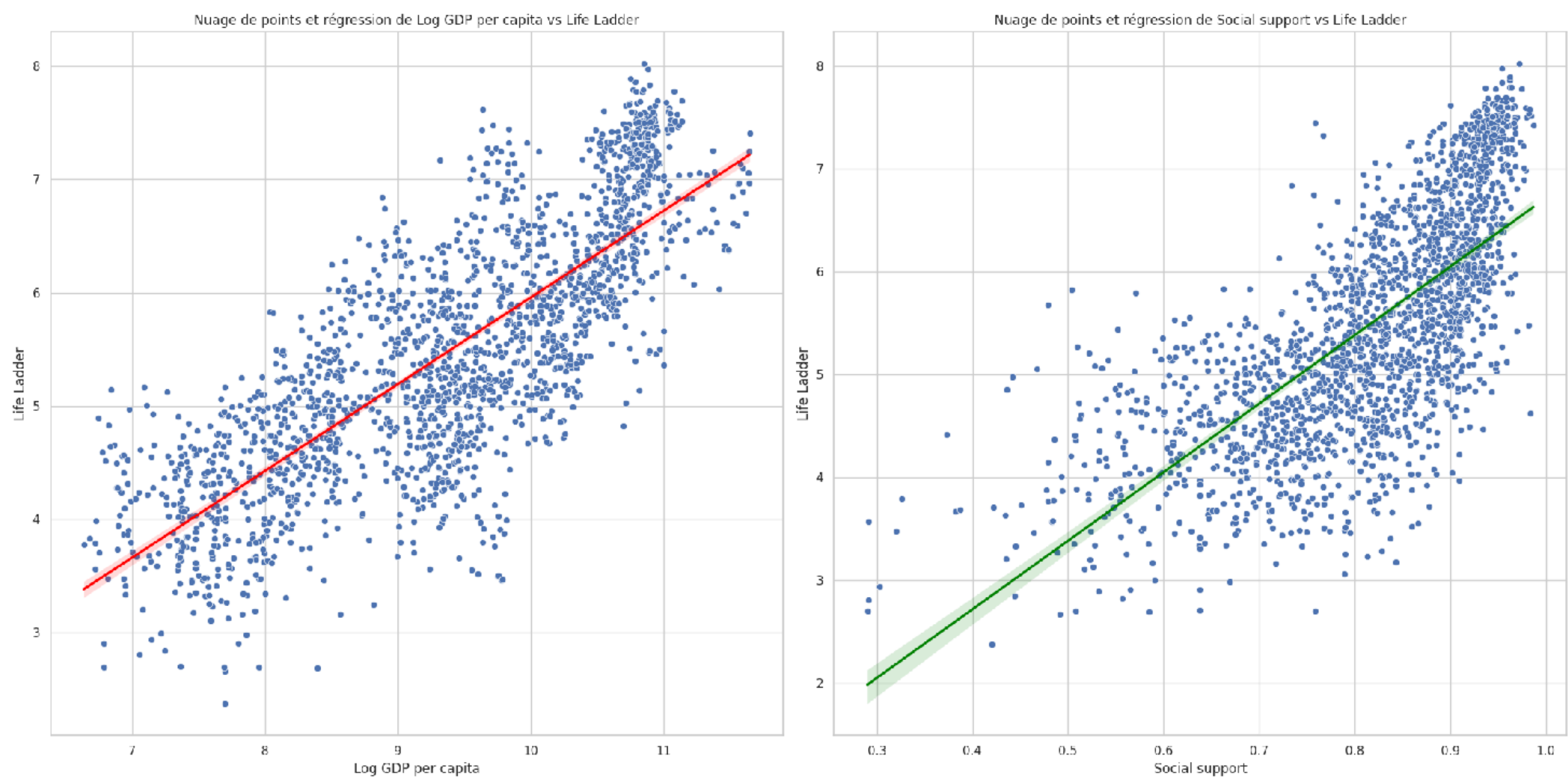
- Le score du bonheur ["Life Ladder"] et l'économie au travers de la variable "Log GDP per capita"
- Le score du bonheur ["Life Ladder"] et la santé au travers de la variable "Life expectancy at birth"
- Le score du bonheur ["Life Ladder"] et la politique au travers de la variable "Freedom to make life choices"
- Le score du bonheur ["Life Ladder"] et le changement climatique au travers de la variable "Température"

Nous avons pris le parti de ne présenter dans ce rapport que les nuages de points et la droite de régression afin de démontrer la relation existante entre chacune des différentes variables listées ci-dessus et le score du bonheur.

Concernant le détail de notre analyse, il est consultable au niveau de la partie “Étape 2/ Exploration et analyse des données avec DataViz” via ce lien : [Travail de groupe - Colab \(google.com\)](#)

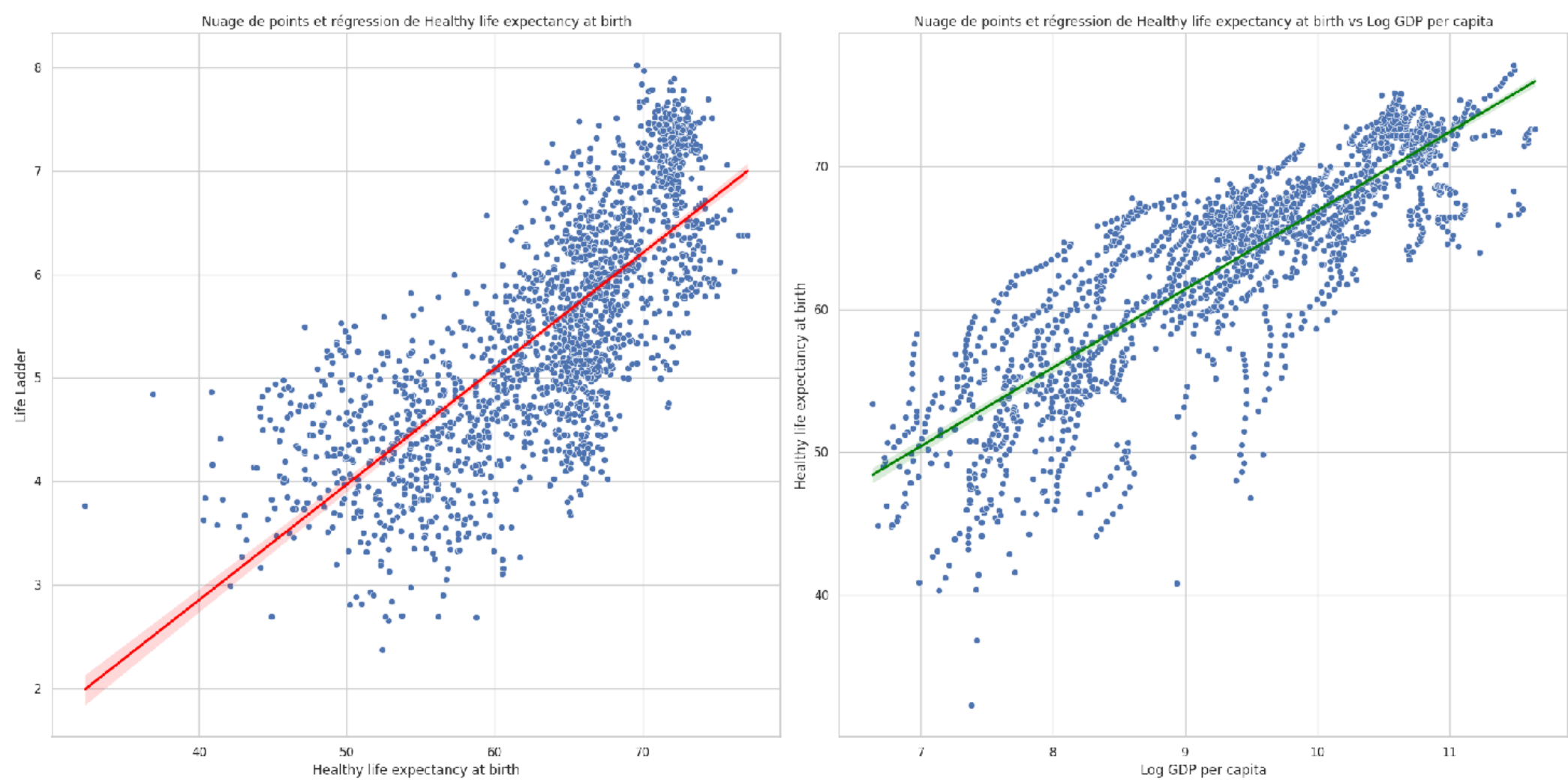
1) Relation entre le score du bonheur et l'économie

Il y a une forte corrélation positive entre les variables PIB par habitant (Log GDP per capita) et Score de bonheur (Life Ladder). Cela suggère que les pays avec un PIB par habitant plus élevé ont tendance à avoir des scores de bonheur plus élevés.



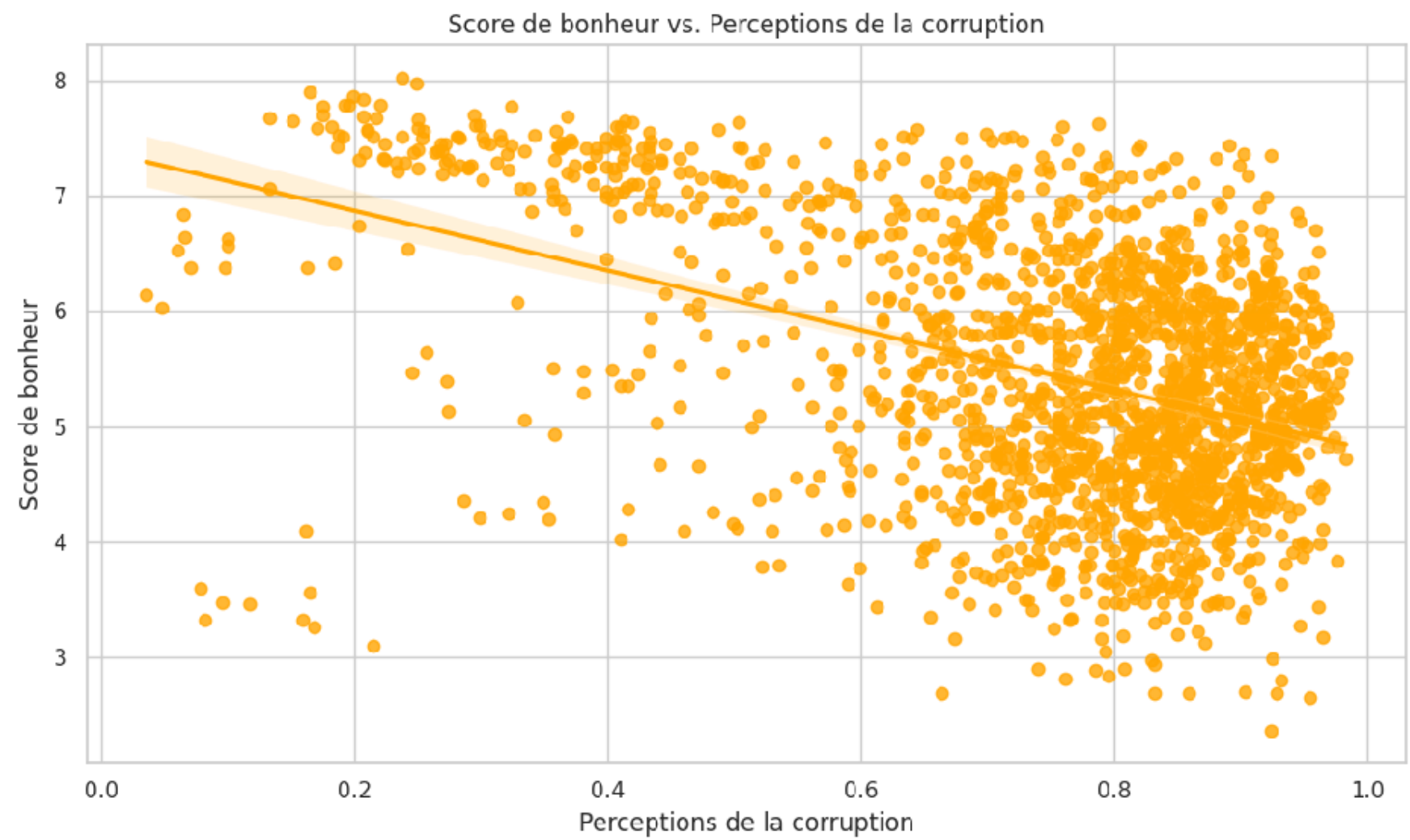
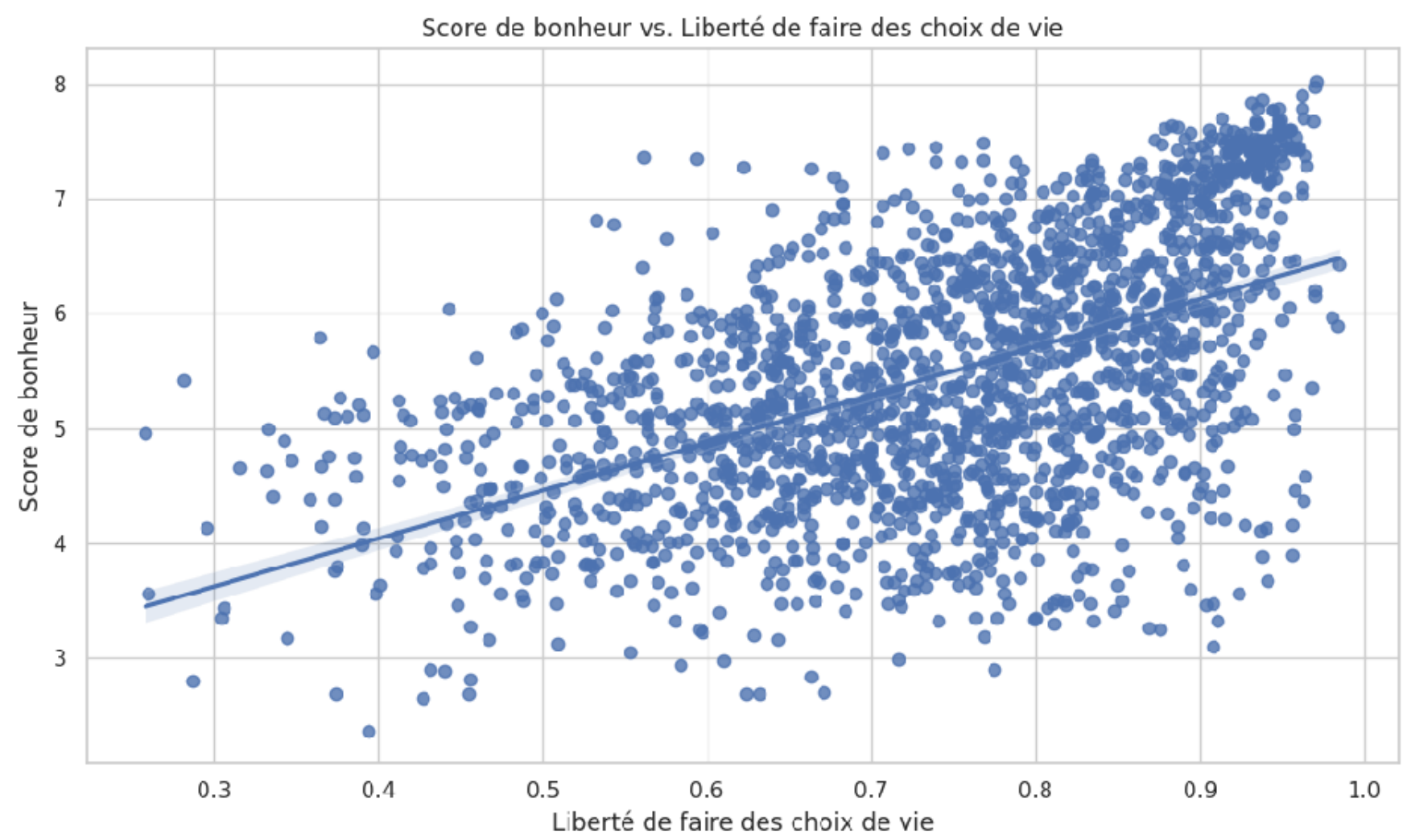
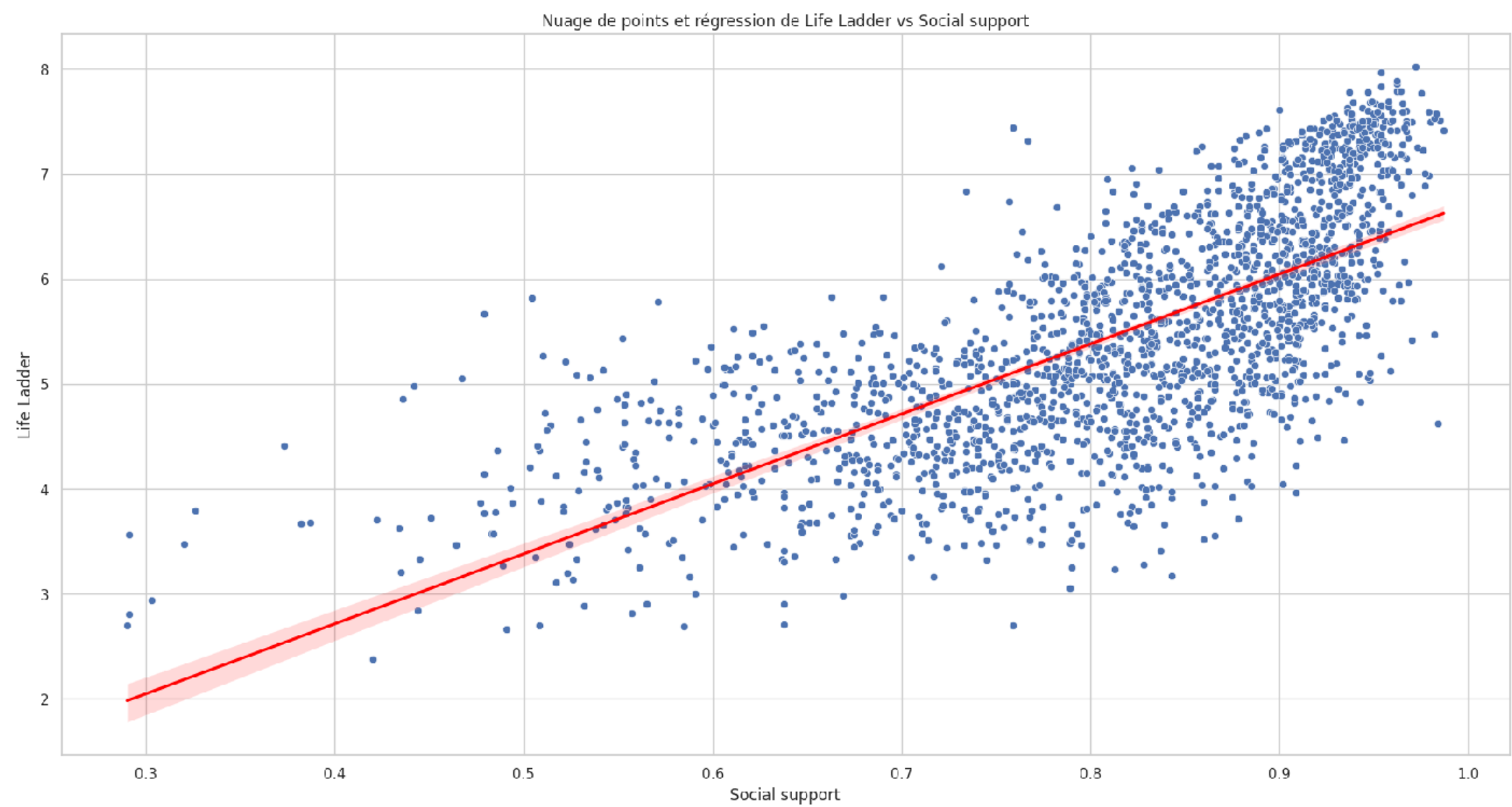
2) Relation entre le score du bonheur et la santé

Il y a une forte corrélation positive entre les variables espérance de vie à la naissance (Healthy life expectancy at birth) et Score de bonheur (Life Ladder). Cela suggère que les pays ayant une espérance de vie à la naissance élevée ont tendance à avoir des scores de bonheur également élevés.

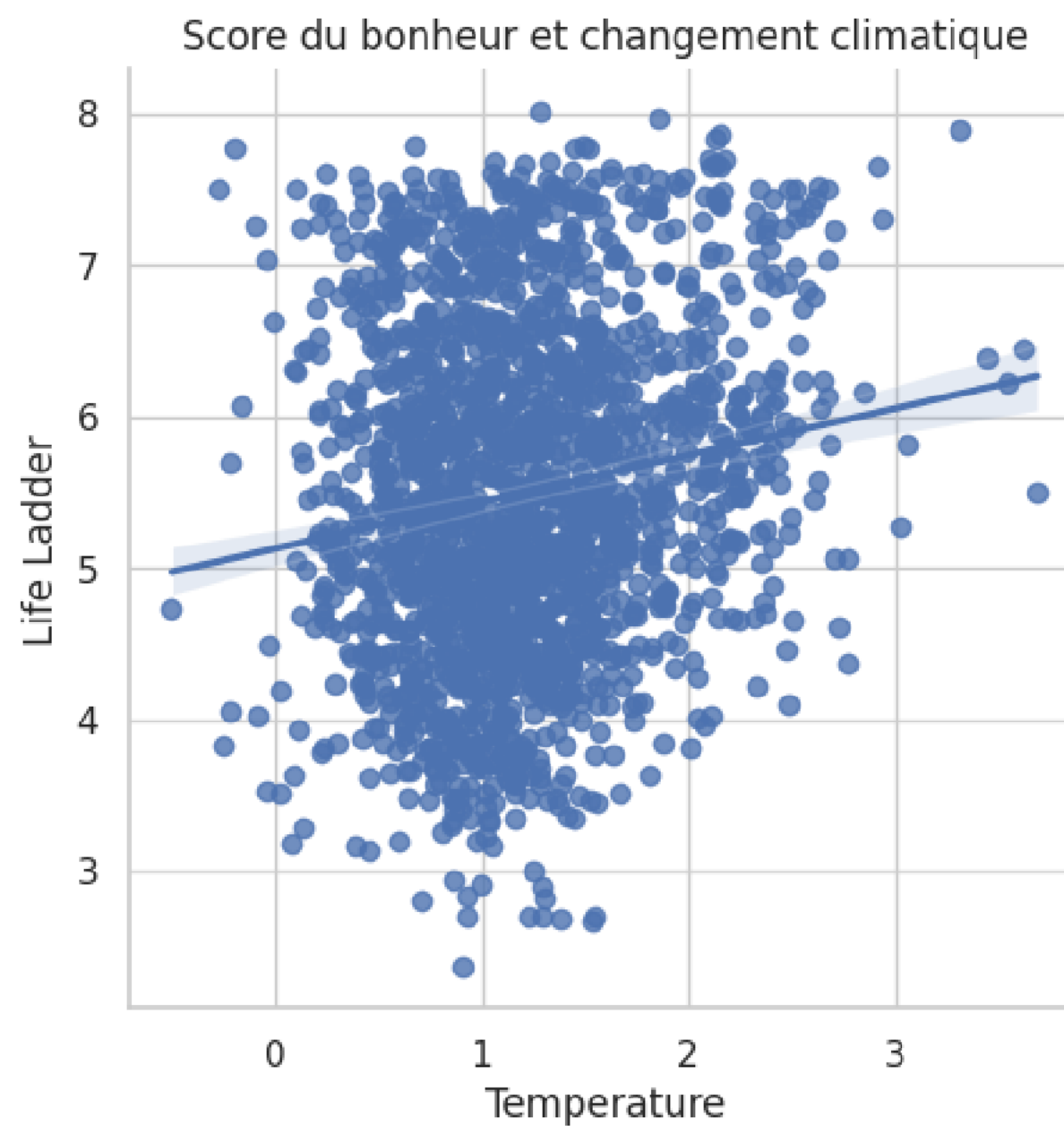


3) Relation entre le score du bonheur et la politique

Il y a également une forte corrélation positive entre le Soutien social (Social support) et Score de bonheur (Life Ladder). Cela semble indiquer que les pays où les individus perçoivent un plus grand soutien social ont tendance à avoir des scores de bonheur plus élevés.



4) Relation entre le score du bonheur et le changement climatique



Il ne semble pas y avoir de corrélation entre le score du bonheur et le changement climatique.

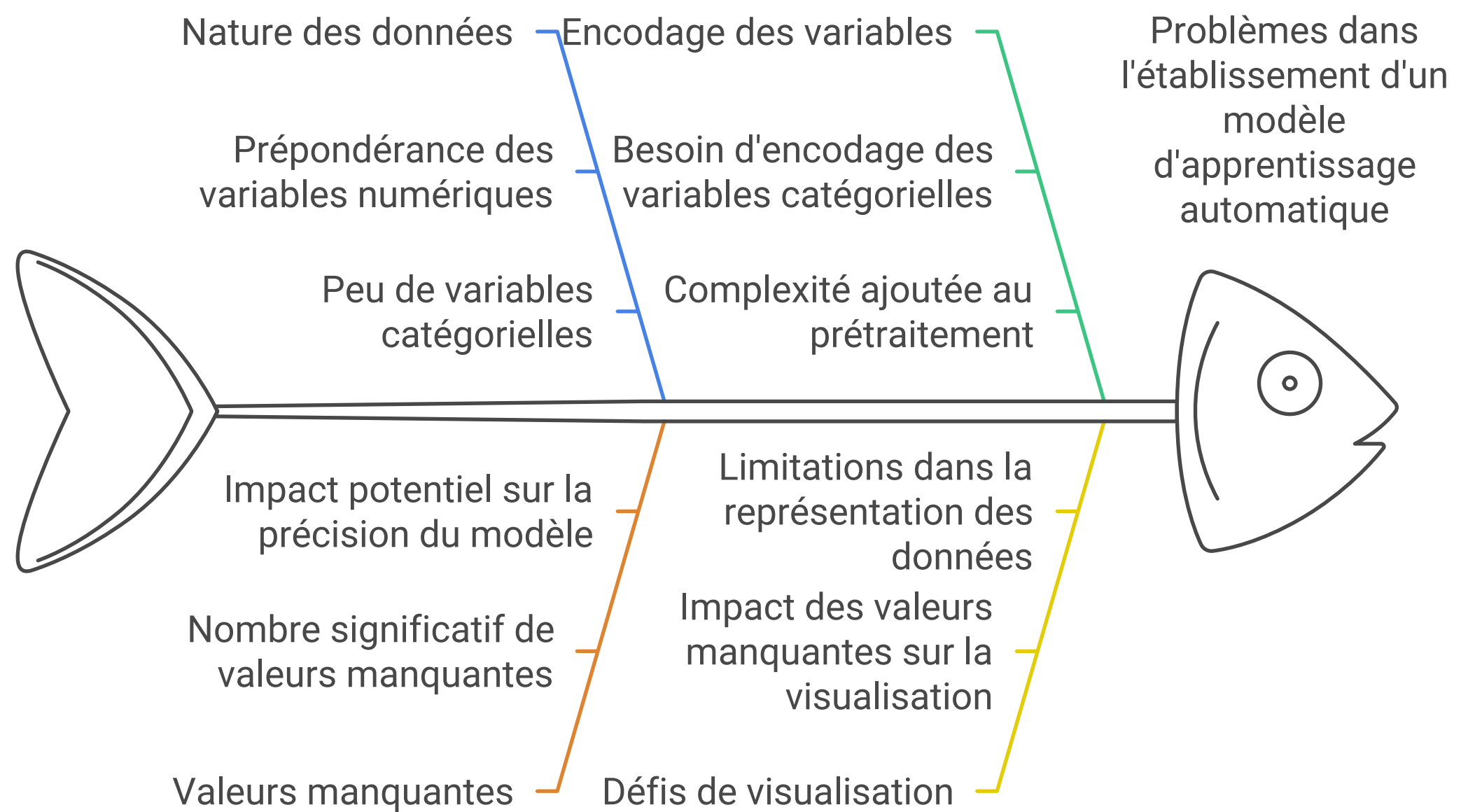
IV. Preprocessing

A) **Constat général sur le jeu de données**

Le jeu de données utilisé est principalement composé de variables numériques. En effet, seuls les colonnes "Country name" et "year" contiennent des variables catégorielles.

Conséquemment, dans le cadre du preprocessing, notre principale action se situera dans l'encodage des variables. Toutefois, pour donner suite à nos travaux de visualisation, nous notons l'existence d'un nombre significatif de valeurs manquantes qui pourraient avoir une incidence sur l'établissement d'un modèle de machine learning supervisé.

Défis dans le prétraitement des données pour l'apprentissage automatique



1) Recherche des valeurs manquantes

Le détail des valeurs manquantes par variables est consultable dans la partie "Etape 1/ comprendre les données" via ce lien : [Travail de groupe - Colab \[google.com\]](#) :

```
df.isna().sum() #affichage des valeurs manquantes pour chaque variable
```

| | |
|----------------------------------|-----|
| | 0 |
| Country name | 0 |
| year | 0 |
| Life Ladder | 0 |
| Log GDP per capita | 36 |
| Social support | 13 |
| Healthy life expectancy at birth | 55 |
| Freedom to make life choices | 32 |
| Generosity | 89 |
| Perceptions of corruption | 110 |
| Positive affect | 22 |
| Negative affect | 16 |

dtype: int64

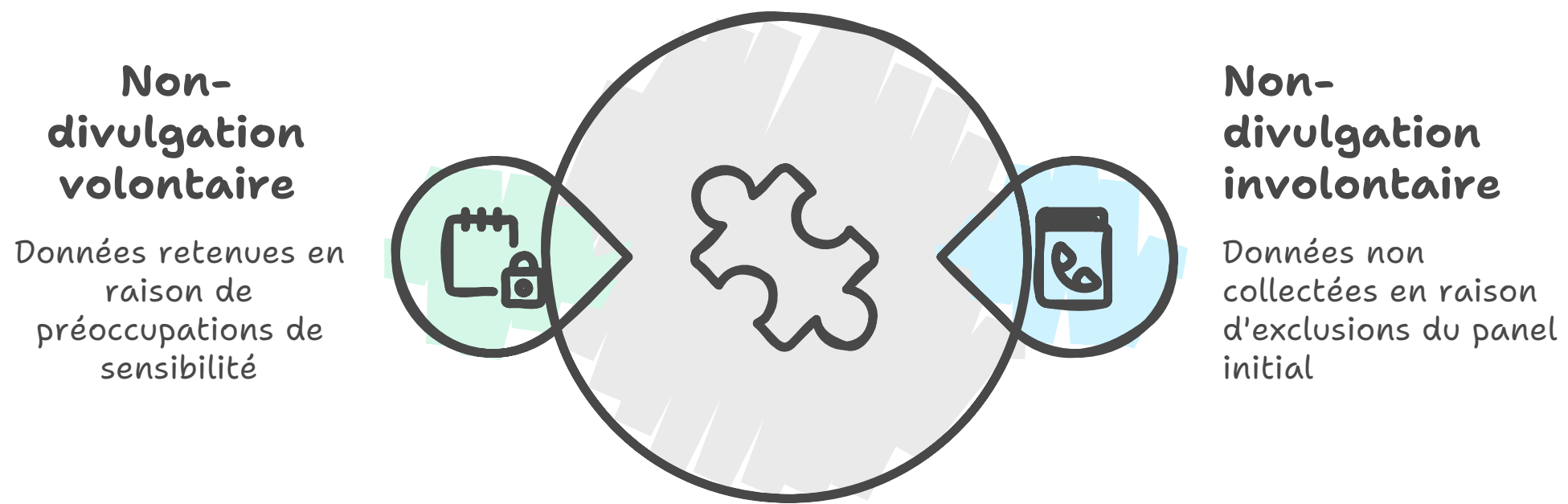
2) Synthèse explicative sur les valeurs manquantes

L'absence de ces données pourrait être liée aux causes suivantes :

- Non-divulgence involontaire des données (collecte de données non réalisée car pays non inclus dans le panel initial par exemple)

- Non-divulgence volontaire des données [décision étatique de ne pas partager des données jugées sensibles]

Causes des données manquantes

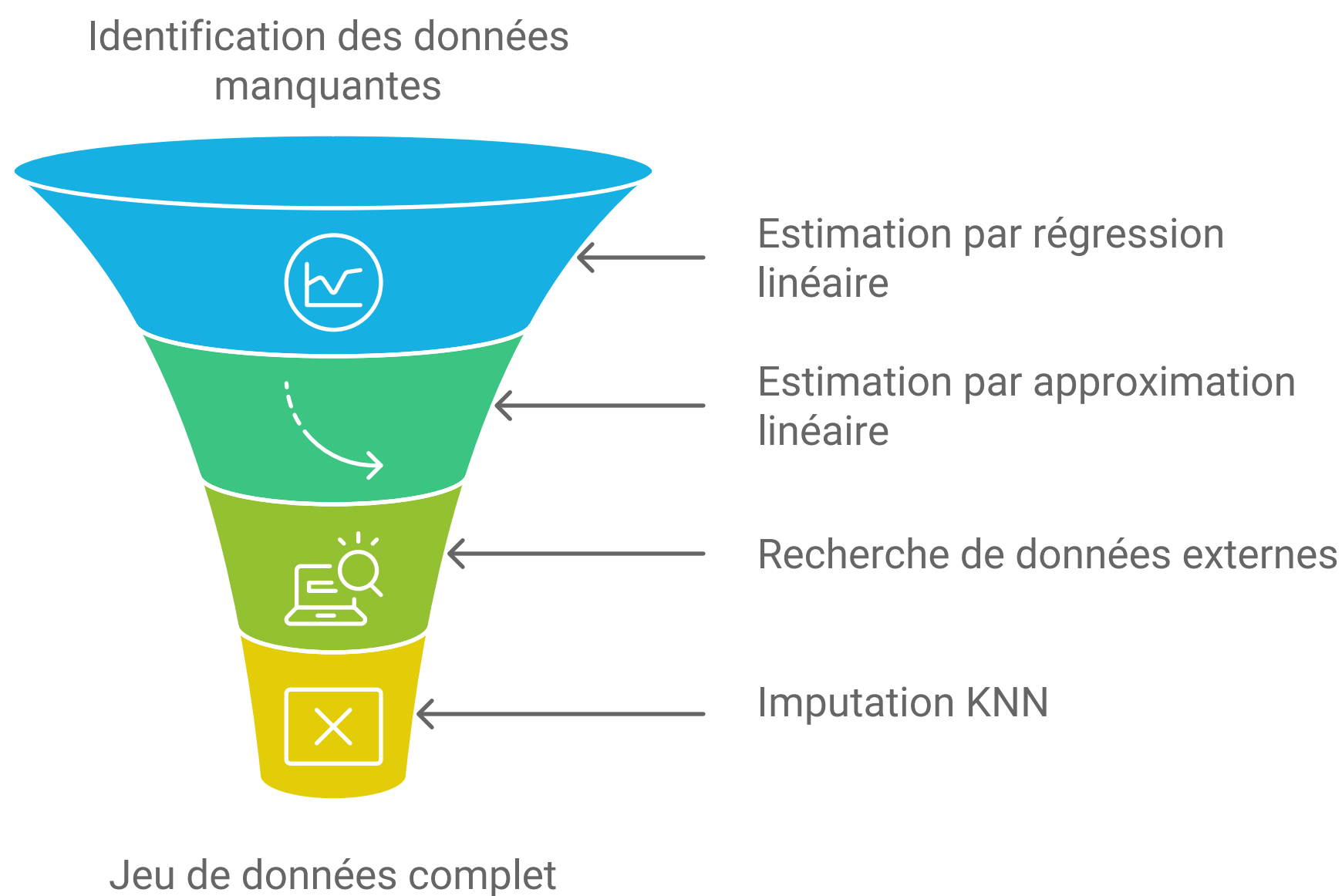


B) Explication de notre démarche pour remplacer les données manquantes

Afin de nettoyer notre jeu de données, nous avons opté pour les démarches suivantes en fonction des cas :

- 1) **Estimation par régression linéaire** dans le cas où des valeurs seraient manquantes pour un pays sur les dernières années (*par exemple : pour l'Australie où il manque les données de la variable "Healthy Life Expectancy at birth" de 2018 à 2021.*)
- 2)
- 3) **Estimation par approximation linéaire** dans le cas où des valeurs seraient manquantes pour un pays sur un intervalle de temps (*par exemple : pour le Cap Vert où il manque les données climatiques pour l'année 2010 et 2017*)
- 4) **Recherche de données externes par scraping ou création de liste puis ajout sur le data frame** pour les pays qui présentent des données manquantes sur toutes les années. Pour mettre en œuvre la démarche explicitée précédemment, nous avons littéralement appliqué, pour chacune des variables de notre jeu de données contenant des valeurs manquantes, les deux étapes suivantes :
 - 1) Imputation par régression linéaire
 - 2) Imputation par KNN (k-nearest neighbors)
 Par le biais de cette méthodologie, nous avons pu traiter l'ensemble des valeurs manquantes de notre jeu de données.

Processus d'imputation des données



C) Encoding des variables

Afin de finaliser notre étape de preprocessing, nous avons instancié :

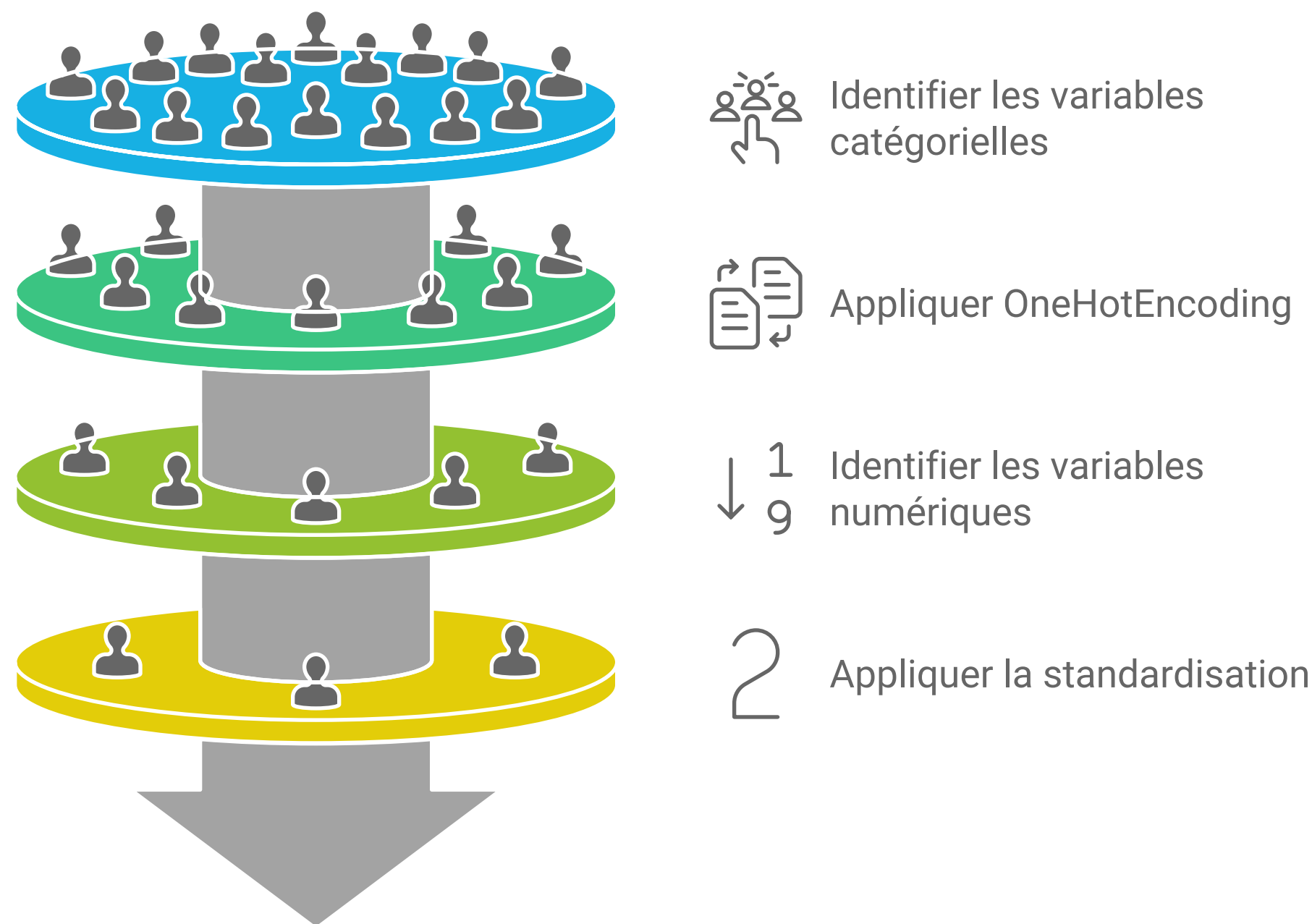
- **Pour les variables catégorielles (i.e “Country name” et “Regional indicator”) :**

§ “OneHotEncoder”

- **Pour les variables quantitatives/numériques :**

§ Une standardisation pour l'ensemble des autres variables de notre dataset

Prétraitement des données pour l'analyse

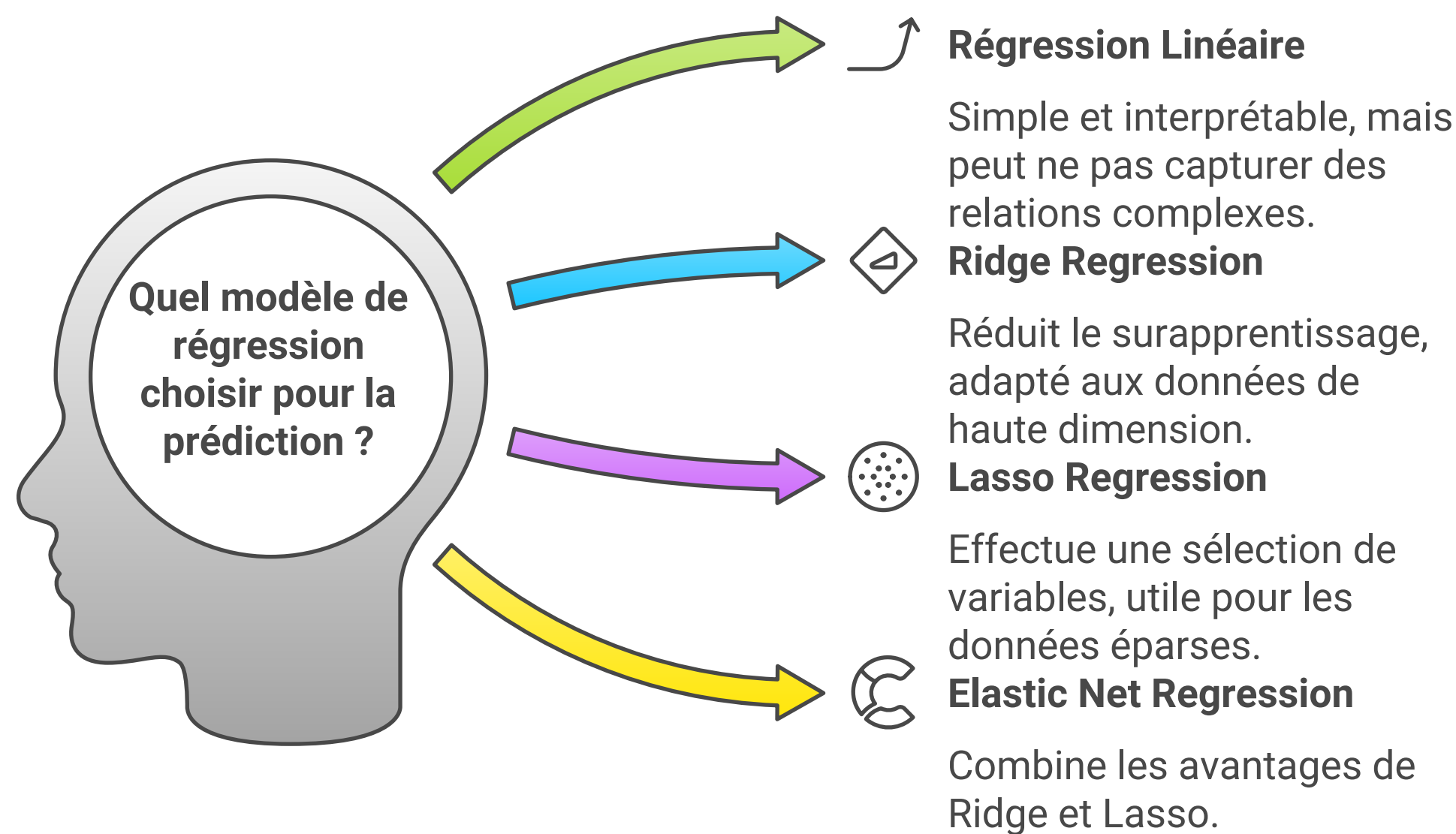


A) Sélection des modèles

L'étape de preprocessing nous a permis de nettoyer l'ensemble du jeu de données ; il n'y a plus de valeur manquante. Cela nous permet de commencer votre phase de modélisation avec un jeu « propre ».

Afin de sélectionner le modèle de prédiction le plus performant pour notre jeu de données, nous avons réalisés une analyse comparative entre les modèles existants suivants :

- Régression Linéaire
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Random Forest Regressor
- XGBoost Regressor

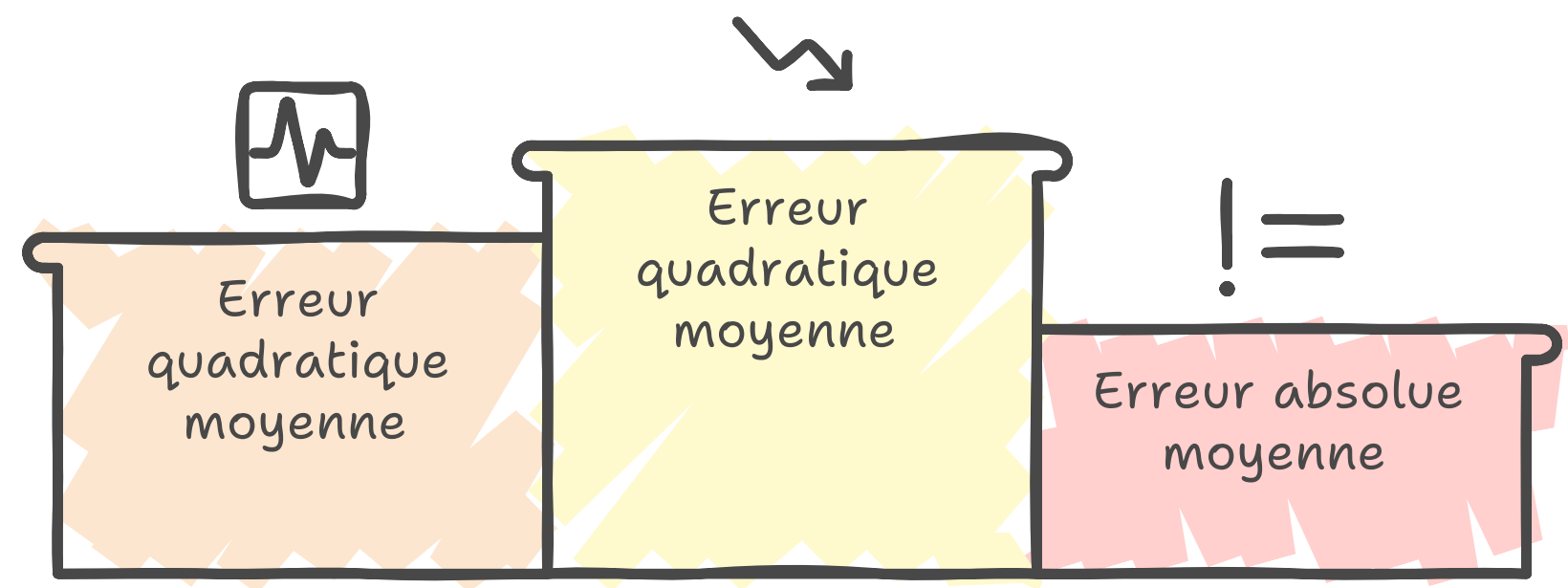


B) Résultat de la modélisation


Dans le cadre de notre analyse, nous avons calculé, pour chacun des modèles listés ci-dessus, les métriques de performance suivantes :

- Mean Absolute Error [MAE]
- Mean Squared Error [MSE]
- Root Mean Squared Error [RMSE]

Métriques de performance dans l'évaluation des modèles



Les résultats obtenus sont résumés dans le tableau ci-dessous :



Résultats sur l'ensemble d'entraînement :

| | MSE | RMSE | R ² |
|-------------------|----------|----------|----------------|
| Linear Regression | 0.107531 | 0.327919 | 0.913183 |
| Ridge Regression | 0.110788 | 0.332848 | 0.910554 |
| Lasso Regression | 0.285114 | 0.533961 | 0.769808 |
| Elastic Net | 0.441869 | 0.664732 | 0.643251 |
| Random Forest | 0.018576 | 0.136295 | 0.985002 |
| XGBoost | 0.036765 | 0.191743 | 0.970317 |

Résultats sur l'ensemble de test :

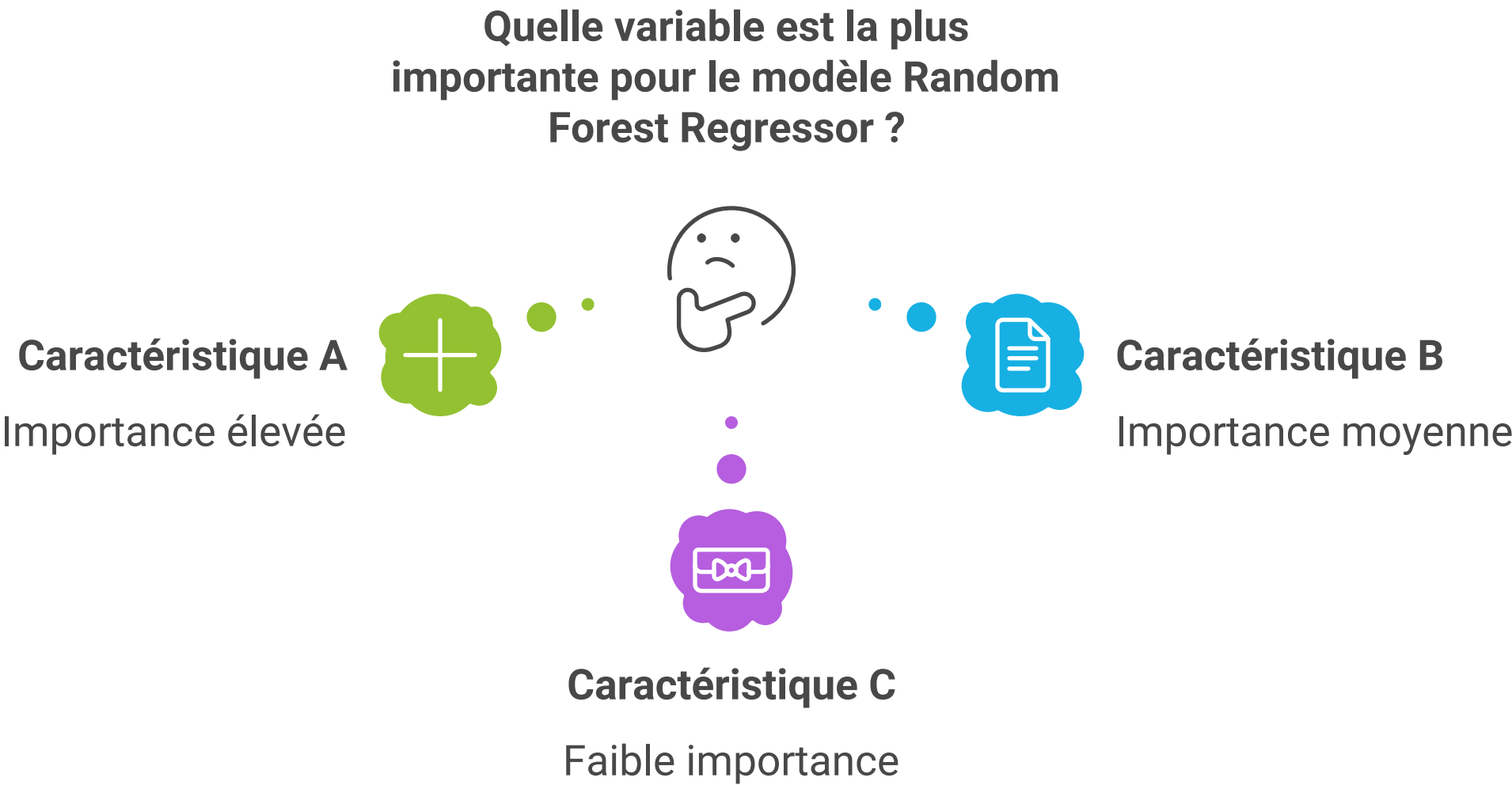
| | MSE | RMSE | R ² |
|-------------------|----------|----------|----------------|
| Linear Regression | 0.148302 | 0.385100 | 0.882887 |
| Ridge Regression | 0.149736 | 0.386957 | 0.881755 |
| Lasso Regression | 0.316851 | 0.562895 | 0.749786 |
| Elastic Net | 0.474315 | 0.688705 | 0.625438 |
| Random Forest | 0.144992 | 0.380779 | 0.885501 |
| XGBoost | 0.142702 | 0.377759 | 0.887310 |

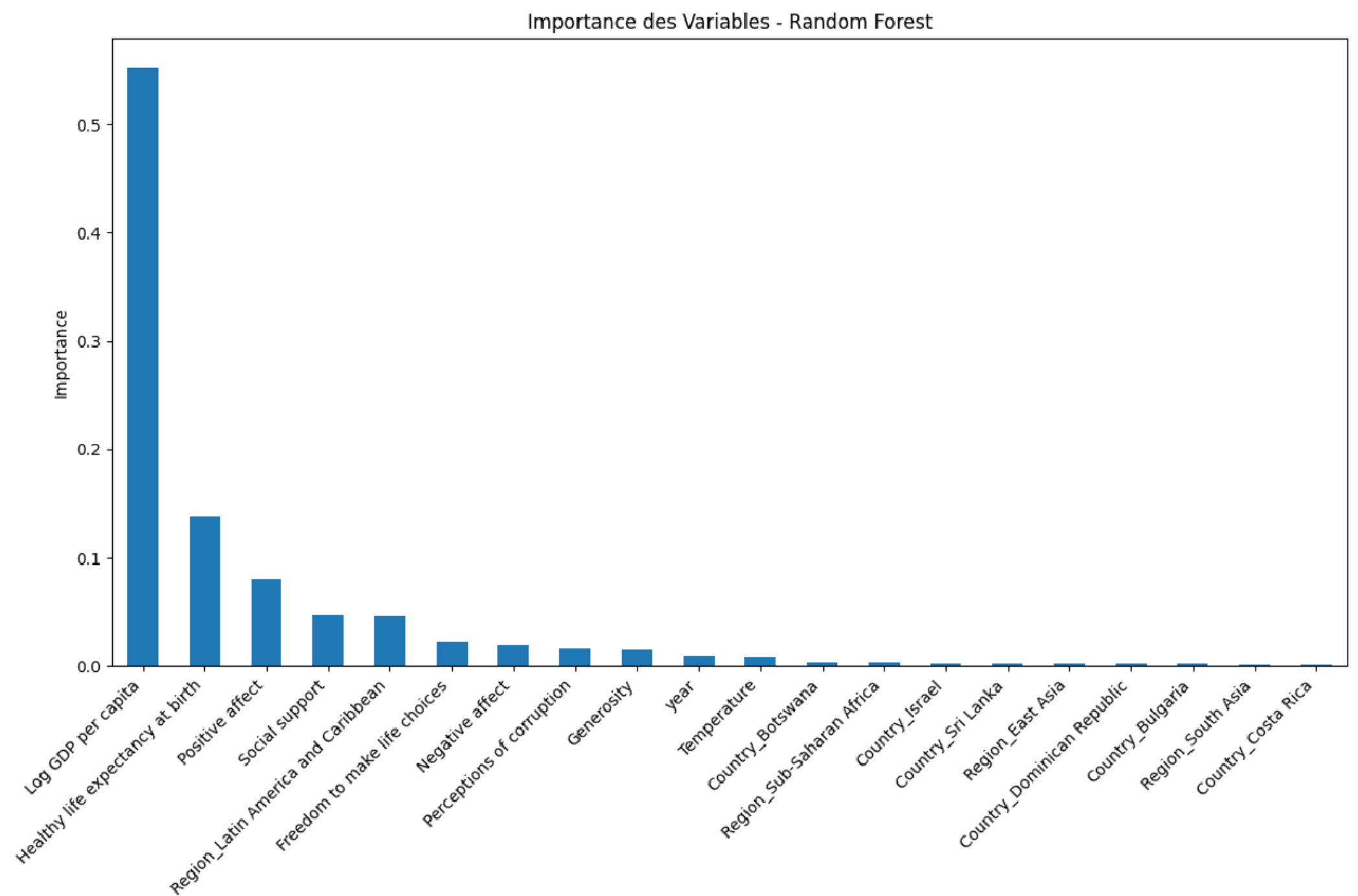
Afin de sélectionner le modèle le plus opérant pour notre étude, nous nous sommes basés principalement sur les résultats du **MSE** et du **RMSE**.
En nous appuyant principalement sur ces deux résultats, nous pouvons conclure que les modèles les plus performants sont le Random **Forest Regressor** puis le **XGBoost**.

Ces résultats montrent que **Random Forest** et **XGBoost** offrent les meilleures performances avec ou sans outliers, tandis que la **régression linéaire** est particulièrement sensible aux outliers, rendant nécessaire leur suppression pour obtenir des résultats raisonnables.

1. Niveau d'importance des variables

Nous avons cherché ensuite à quantifier, pour le modèle le plus performant : i.e Random Forest Regressor, le rôle de chacune des variables en étudiant les “features importances” afin de classer les variables selon leur niveau d’implication dans les choix de prédiction.
Grâce à l’attribut *feature_importances* nous avons obtenu les résultats suivants pour le modèle Random Forest Regressor :





Les features importances nous indiquent les quatres variables les plus importantes :

1. Log GDP per capita
2. Healthy life expectancy at birth
3. Positive affect
4. Social support

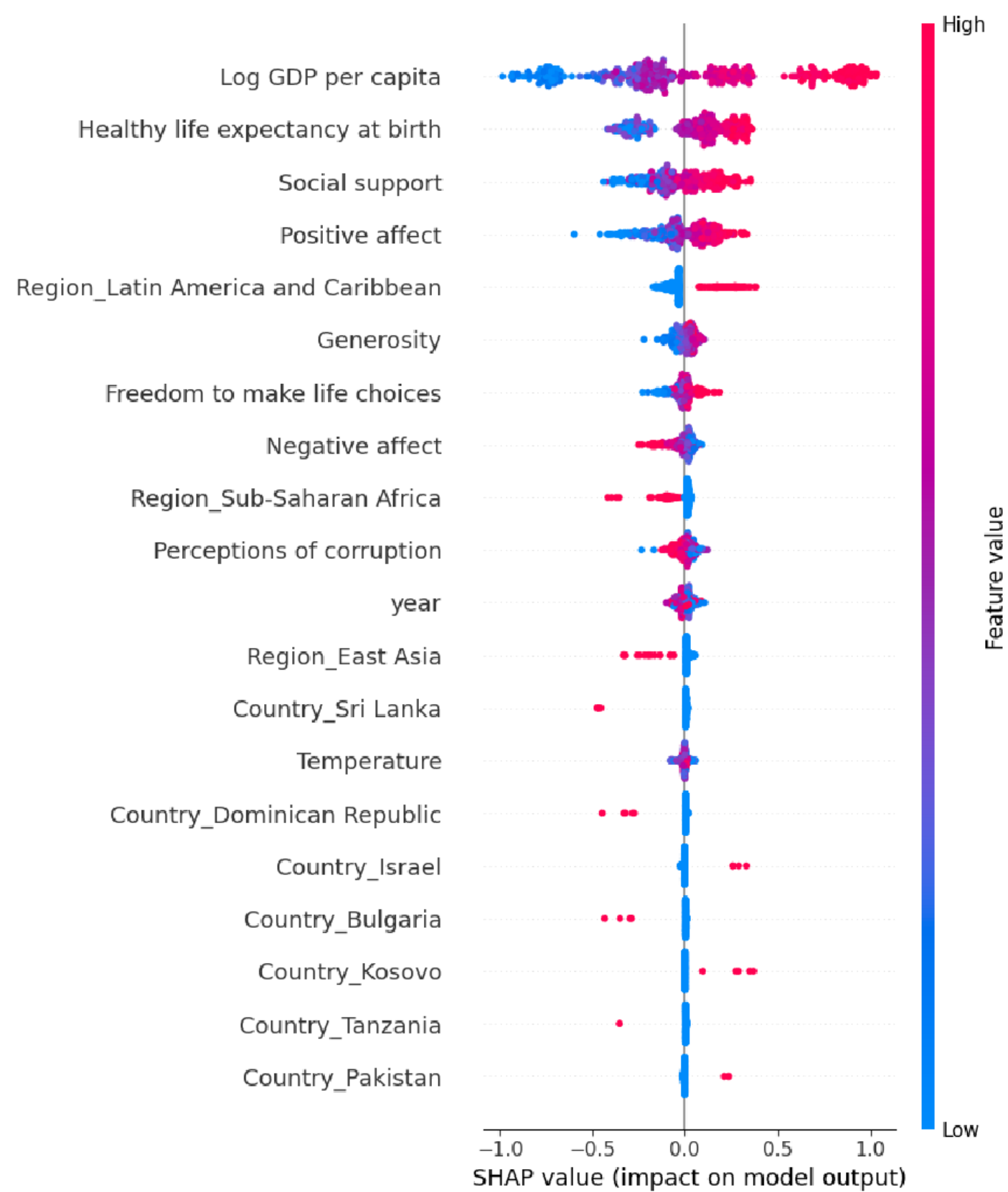
1. Interprétabilité des modèles

Pour comprendre l'importance des variables et leur impact sur nos prédictions, nous avons utilisé SHAP.

- Pour le modèle Random Forest Regressor :



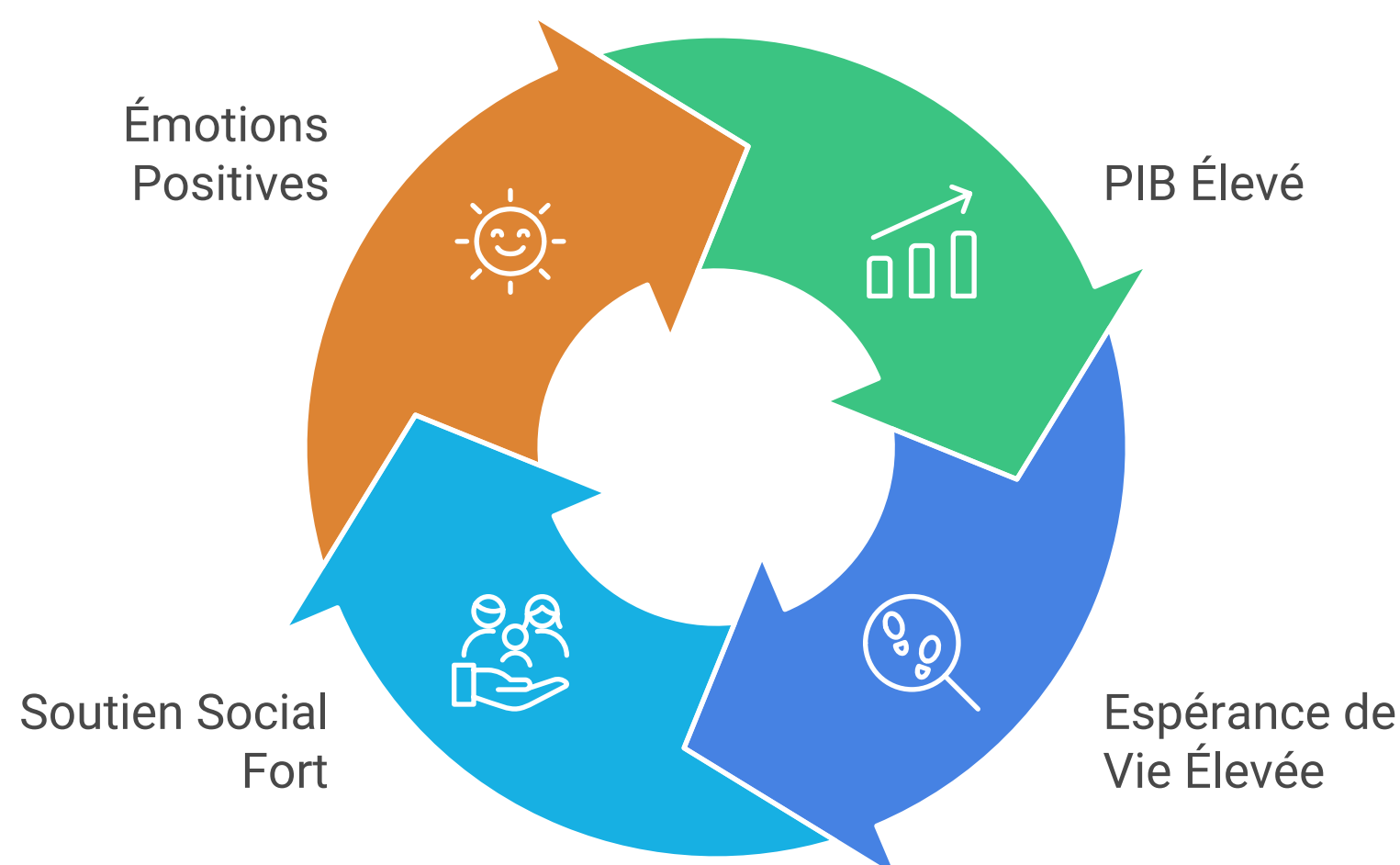
- Pour le modèle XGBoost :



Dans les deux cas de figure le constat est le suivant :

- Un niveau de PIB élevé augmente la corrélation avec le score du bonheur
- Un niveau élevé d'espérance de vie à la naissance augmente la corrélation avec le score du bonheur
- Un niveau élevé de support social augmente la corrélation avec le score du bonheur
- Un niveau élevé de sentiments dits positifs augmente la corrélation avec le score du bonheur

Cycle de Corrélation du Bonheur



Entre autres, les variables explicatives jouant un rôle important dans la prédiction du score du bonheur pour nos deux modèles contribuent à l'augmentation du score du bonheur si elles sont élevées et inversement.

1. Conclusion

Bien que notre étude ait fourni des résultats significatifs, elle ouvre également la voie à de futures recherches. Il serait intéressant d'explorer l'impact d'autres variables telles que le taux de chômage, la croissance économique, ou encore l'inflation et l'éducation, afin d'affiner nos prédictions.

De plus, une estimation du score de bonheur pour l'année 2021, en comparaison avec les données réelles, pourrait offrir une validation supplémentaire de nos modèles.

Enfin, l'intégration de modèles plus sophistiqués ou l'optimisation des hyperparamètres des modèles existants pourrait permettre d'améliorer encore les performances de nos prédictions.

En somme, cette étude offre non seulement une analyse approfondie des facteurs influençant le bonheur à travers le monde, mais elle constitue également une base solide pour des recherches futures qui pourraient aider les décideurs politiques à orienter leurs stratégies en faveur du bien-être de leurs citoyens.

Perspectives et améliorations

- Explorer l'impact d'autres variables pour affiner l'étude (taux de chômage, inflation, éducation).
- Estimer le score du bonheur pour l'année 2021 et le comparer aux données réelles.
- Optimiser les hyperparamètres des modèles existants pour améliorer les performances de prédiction.
- Intégrer des modèles plus sophistiqués.

Intérêts de l'étude

- Analyse globale des facteurs influençant le bonheur mondial.
- Base pour de futures recherches.
- Eventuel support d'orientation de stratégies politiques.

