SAE – Analyse statistique du bonheur dans le monde

Quels sont les facteurs qui influencent le plus le bonheur d'un pays ?

10/06/2025

Yacine BABOURI, Rida AIT-ZAOUIT, Younes CHIKER,
Groupe 11-12
BUT 1 Science des données

Sommaire

Problématique :

Quels sont les facteurs qui influencent le plus le bonheur d'un pays ?

I. Introduction

Ce projet vise à identifier les variables ayant un impact significatif sur le score de bonheur. Grâce à l'analyse des données du **World Happiness Report**, nous tenterons de comprendre quels éléments — parmi le PIB par habitant, le soutien social, la liberté de choix ou encore la perception de la corruption — jouent un rôle déterminant dans le niveau de bonheur ressenti par les populations.

Le World Happiness Report est une étude publiée chaque année qui analyse et classe les pays en fonction du niveau de bonheur perçu par leurs habitants. Depuis 2012, ce rapport évalue le bien-être de plus de 150 pays à travers des enquêtes à grande échelle. Il est produit par le **Sustainable Development Solutions Network** (SDSN) des Nations Unies et repose sur des données collectées directement auprès des citoyens, qui évaluent leur propre bonheur sur une échelle de 0 à 10.

En complément de ces données subjectives, plusieurs indicateurs socio-économiques et politiques sont intégrés afin d'expliquer les différences de perception du bonheur entre les pays. Ces indicateurs incluent notamment le **PIB par habitant**, le **soutien social**, l'**espérance de vie**, la **liberté individuelle**, la **générosité** et la **perception de la corruption**.

Le rapport ne se limite pas à un simple classement : il analyse les tendances mondiales, met en lumière l'évolution du bonheur dans le temps et explore les causes des écarts observés. L'objectif est d'identifier les déterminants du bien-être afin d'aider les décideurs publics à élaborer des politiques visant à améliorer la qualité de vie des populations.

En 2024, le World Happiness Report couvre 143 pays. Les analyses reposent sur des variables quantitatives permettant d'établir des comparaisons objectives entre les pays. Ces indicateurs facilitent une meilleure compréhension des mécanismes influençant le bien-être et permettent d'expliquer les différences de satisfaction de vie entre les sociétés.

Dans le cadre du **premier rapport**, nous avons :

- Sélectionné la base de données du *World Happiness Report 2024*
- Converti les fichiers Excel en format .csv puis .xls (.csv ne fonctionnait pas) pour une meilleure compatibilité avec R
- Traduit et clarifié les intitulés des variables
- Nettoyé les données manuellement à l'aide d'Excel
- Réalisé une première exploration des données : moyennes, corrélations visuelles, graphiques descriptifs (diagrammes en barres, nuages de points)

Le présent rapport reprend ce travail et l'approfondit, en y intégrant une analyse statistique plus rigoureuse et une documentation complète des traitements réalisés avec le logiciel R.

Objectif de ce rapport

L'objectif de ce rapport est d'approfondir l'analyse amorcée dans la première phase du projet en exploitant les données du World Happiness Report à l'aide du logiciel R.

Plus précisément, il s'agit de :

- Réaliser des statistiques descriptives et exploratoires
- Identifier les corrélations entre les différentes variables explicatives
- Estimer l'influence de chaque facteur sur le bonheur à l'aide de régressions linéaires
- Visualiser les relations observées à l'aide de graphiques pertinents

II. Description des données brutes

Nous avons choisi de nous concentrer exclusivement sur le fichier **Data_WHR2024.xls**, comme annoncé dans notre premier rapport. Ce fichier contient les données les plus récentes disponibles sur le bonheur à l'échelle mondiale. En nous appuyant uniquement sur cette version, nous visons une analyse plus **pertinente et actuelle**, tout en **évitant les biais temporels** potentiellement introduits par les données issues des éditions précédentes du rapport (différences de méthodologie, changement de pays inclus, évolution des indicateurs, etc.).

Les données exploitées proviennent du **World Happiness Report 2024**, lui-même fondé sur les résultats du **Gallup World Poll**, une enquête internationale de référence sur le bien-être subjectif des populations. La version utilisée ici a été préalablement **nettoyée et traduite**, afin de faciliter son exploitation statistique tout en assurant la cohérence des noms de variables.

Dans cette base de données, **chaque ligne correspond à un pays**, et les **colonnes** représentent les principaux indicateurs utilisés pour expliquer ou illustrer le niveau de bonheur déclaré par les habitants. Les variables principales sont :

- **Score** : score moyen de bonheur, compris entre 0 et 10, tel que déclaré par les individus dans chaque pays ;
- **PIB/Habit** : produit intérieur brut par habitant, exprimé en logarithme pour lisser les disparités extrêmes ;
- **Soutien Social** : indicateur mesurant la perception qu'ont les individus de pouvoir compter sur des proches en cas de besoin ;
- **Espérance de Vie** : espérance de vie en bonne santé à la naissance ;
- **Liberté** : degré de liberté perçue par les individus pour faire des choix dans leur vie ;
- **Générosité** : niveau de générosité, basé sur les dons et l'aide à autrui dans le cadre de l'enquête ;
- **Corruption** : perception de la corruption au sein du gouvernement et des entreprises.

Ces variables constituent le socle de notre analyse exploratoire et comparative, avec pour objectif de mieux comprendre les déterminants du bonheur dans le monde en 2024.

III. Nettoyage et création de la base propre

Les données brutes ont été importées dans R à l'aide de la fonction ***read_excel()***, qui permet de lire directement les fichiers au format Excel (.xlsx). Une fois les données chargées, une première phase de nettoyage a été réalisée afin de garantir la qualité et la fiabilité des analyses statistiques futures.

Tout d'abord, nous avons procédé à une inspection générale du jeu de données pour repérer les éventuelles valeurs manquantes ou incohérentes. Afin de limiter l'impact des données incomplètes sur les résultats, nous avons appliqué un critère de suppression : toutes les lignes (observations) contenant plus de deux valeurs manquantes ont été retirées de la base. Ce seuil a été choisi pour maintenir un bon compromis entre la conservation d'un volume de données suffisant et la fiabilité des analyses.

Cette étape de filtrage a permis de constituer une base de données plus propre, contenant des observations suffisamment complètes pour que les résultats des traitements statistiques ne soient pas biaisés par une trop grande quantité de données absentes. Le reste des valeurs manquantes (moins de trois par ligne) a été traité par des méthodes appropriées selon les variables concernées (suppression ponctuelle, imputation, ou exclusion temporaire selon les analyses). Ce nettoyage constitue une étape essentielle dans tout projet d'analyse de données afin d'assurer la robustesse des conclusions tirées par la suite.

IV. Analyse statistique

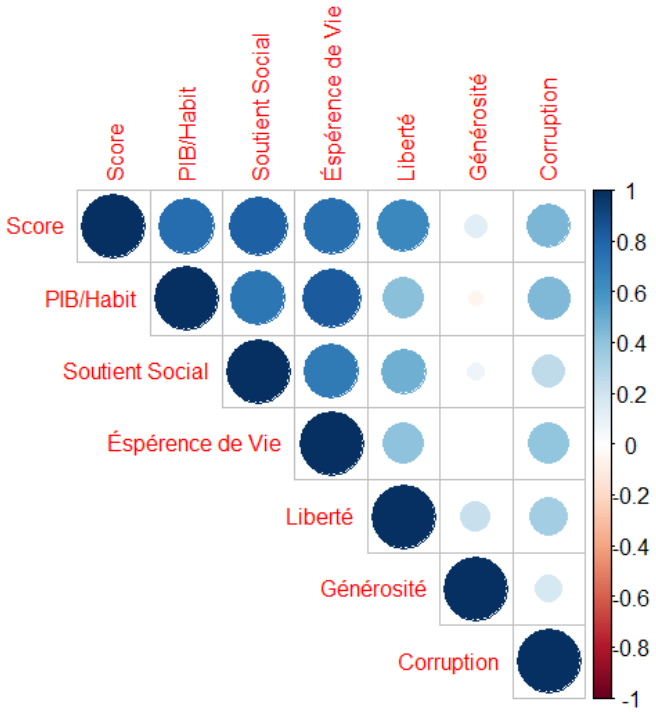
1. Statistiques descriptives

- Le score de bonheur varie entre 1,7 et 7,8
- Moyenne du score : environ 5,5
- Les facteurs comme le soutien social, la liberté et l'espérance de vie présentent des valeurs équilibrées et centrées

2. Corrélation entre variables

Matrice de corrélation entre les variables explicatives et le score de bonheur. Plus la couleur est foncée, plus la corrélation est forte. On observe notamment une forte corrélation entre le score de bonheur et le PIB par habitant, le soutien social ou encore l'espérance de vie.

Matrice de corrélation



Analyse : La matrice de corrélation permet de visualiser les relations linéaires entre les différentes variables explicatives et le score de bonheur. Plus les points sont foncés et proches de 1 (ou de -1), plus la corrélation est forte. L'interprétation de cette matrice met en évidence plusieurs points clés :

- **PIB par habitant** : On observe une forte corrélation positive (environ 0.8) entre le PIB/habitant et le score de bonheur. Cela signifie que, de manière générale, les pays où le revenu moyen par habitant est plus élevé tendent à obtenir des scores de bonheur plus importants. Cela peut s'expliquer par un meilleur accès aux services, à l'éducation, aux soins de santé, et à une qualité de vie matérielle plus confortable.
- **Soutien social, espérance de vie et liberté** : Ces trois variables présentent également des corrélations significatives et positives avec le score de bonheur.
    - Le **soutien social** montre une corrélation élevée, ce qui suggère que la capacité à pouvoir compter sur des proches en cas de besoin joue un rôle majeur dans le bien-être ressenti par les individus.
    - L'**espérance de vie** traduit souvent le niveau de développement sanitaire et la stabilité d'un pays, des éléments qui contribuent au bien-être global.
    - La **liberté** de faire des choix de vie est également un facteur fortement lié au bonheur, traduisant l'importance du sentiment d'autonomie et de liberté individuelle dans l'épanouissement personnel.
- **Générosité et perception de la corruption** : Ces deux variables présentent des corrélations beaucoup plus faibles, voire proches de zéro, avec le score de bonheur.
    - La **générosité** est faiblement corrélée, ce qui peut s'expliquer par des différences culturelles dans la manière dont la générosité est perçue ou mesurée à l'échelle nationale.

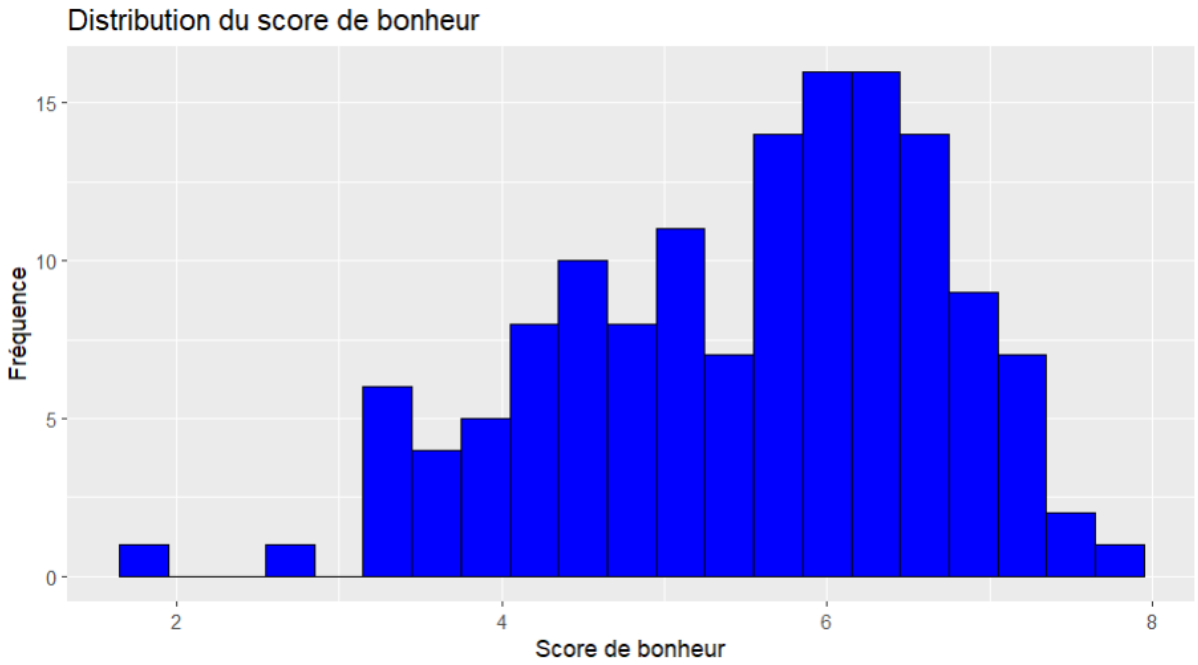
- La **perception de la corruption**, bien qu'importante dans le débat public, n'apparaît pas ici comme fortement liée au score de bonheur dans la matrice de corrélation. Cela peut suggérer que ce facteur influence moins directement le bien-être personnel, ou qu'il interagit avec d'autres variables plus déterminantes.

3. Régression linéaire simple

Code R :

```
modele_simple <- lm(Score ~ `PIB/Habit`, data = df_clean)

summary(modele_simple)
```



Ce script permet de créer un modèle de régression linéaire simple pour analyser si le PIB par habitant a une influence sur le score de bonheur des pays. Le modèle cherche à établir une relation entre ces deux variables. La commande lm() sert à construire ce modèle, et la commande summary() permet d'obtenir un résumé des résultats.

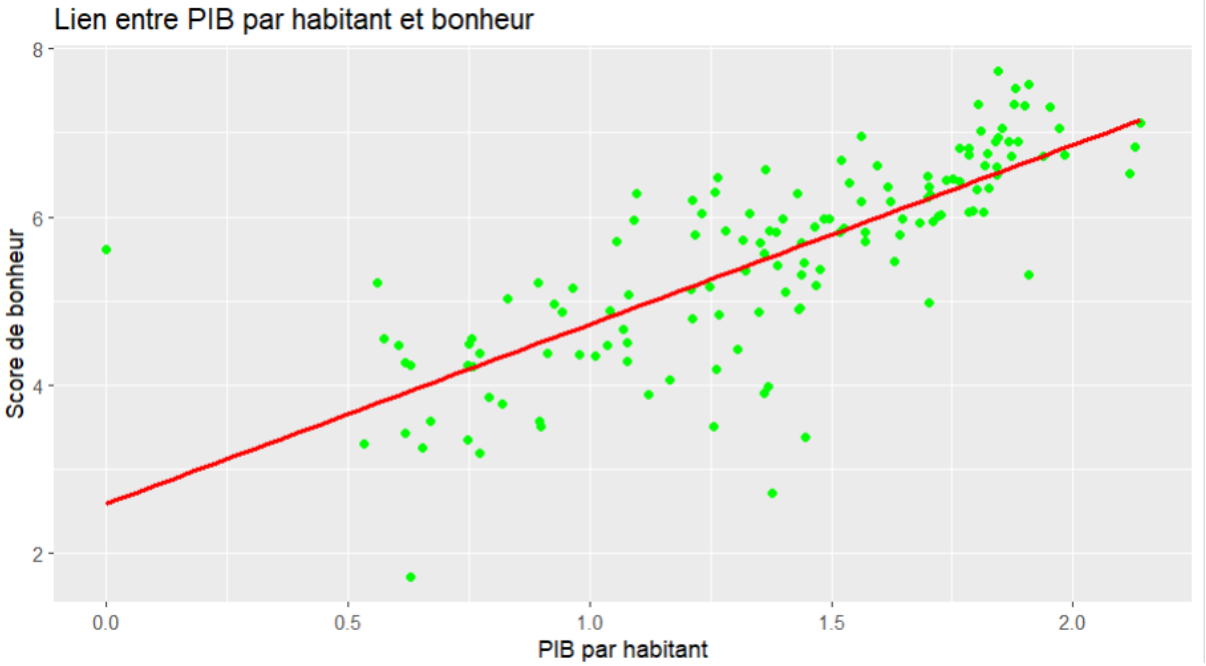
Grâce à ce modèle, on peut voir :

- **si le lien entre le PIB et le bonheur est positif ou négatif,**
- **si ce lien est statistiquement significatif,**
- **et si le PIB explique bien les différences de bonheur entre les pays (grâce au R^2).**

En résumé, ce code sert à vérifier si les pays plus riches ont tendance à être plus heureux.

Par exemple, si le coefficient du PIB est d'environ 1.2, cela signifie qu'une augmentation d'une unité de PIB (en log) est associée à une augmentation moyenne de 1.2 points du score de bonheur.

Si le R^2 du modèle est de 0.65, cela veut dire que 65 % de la variation du score de bonheur entre les pays est expliquée par le PIB.



4. Régression multiple

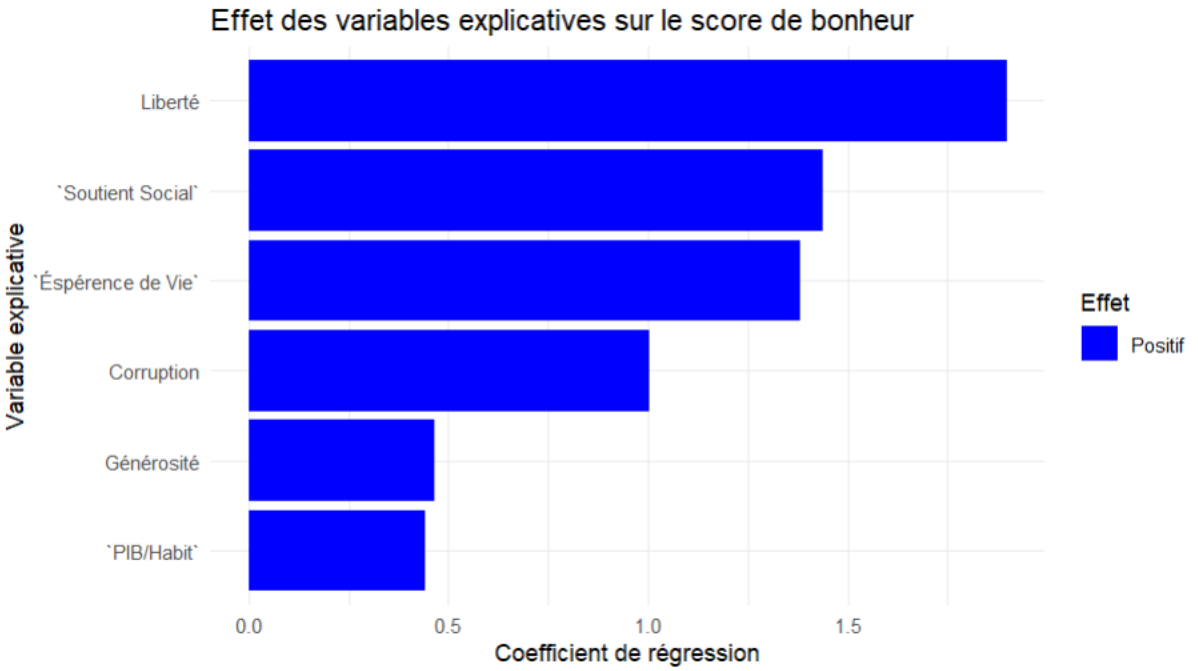
Code R :

```
modele_multiple <- lm(Score ~ `PIB/Habit` + `Soutient Social` + `Éspérence de Vie` + Liberté + Générosité + Corruption, data = df_clean)

summary(modele_multiple)
```

Ce script sert à construire un modèle de régression multiple, c'est-à-dire un modèle qui étudie l'effet de plusieurs variables en même temps sur le score de bonheur. Ici, on analyse l'influence du PIB, du soutien social, de l'espérance de vie, de la liberté, de la générosité et de la corruption.

La commande lm() permet de créer ce modèle, et summary() donne un résumé des résultats.



Grâce à cette régression multiple, on peut :

- **identifier les facteurs les plus importants pour expliquer le bonheur,**
- **voir quelles variables ont un effet significatif ou non,**
- **et mesurer la qualité globale du modèle (avec le R^2).**

Par exemple, le modèle peut montrer que le PIB, le soutien social et la liberté ont des coefficients positifs et significatifs, ce qui indique qu'ils ont une réelle influence sur le bonheur.

Si le R^2 global est de 0.78, cela signifie que 78 % de la variation du score de bonheur est expliquée par l'ensemble des variables prises en compte dans ce modèle.

Ce type de modèle est plus complet que la régression simple car il prend en compte plusieurs aspects du bien-être en même temps.

5. Visualisations

a) Histogramme du score de bonheur

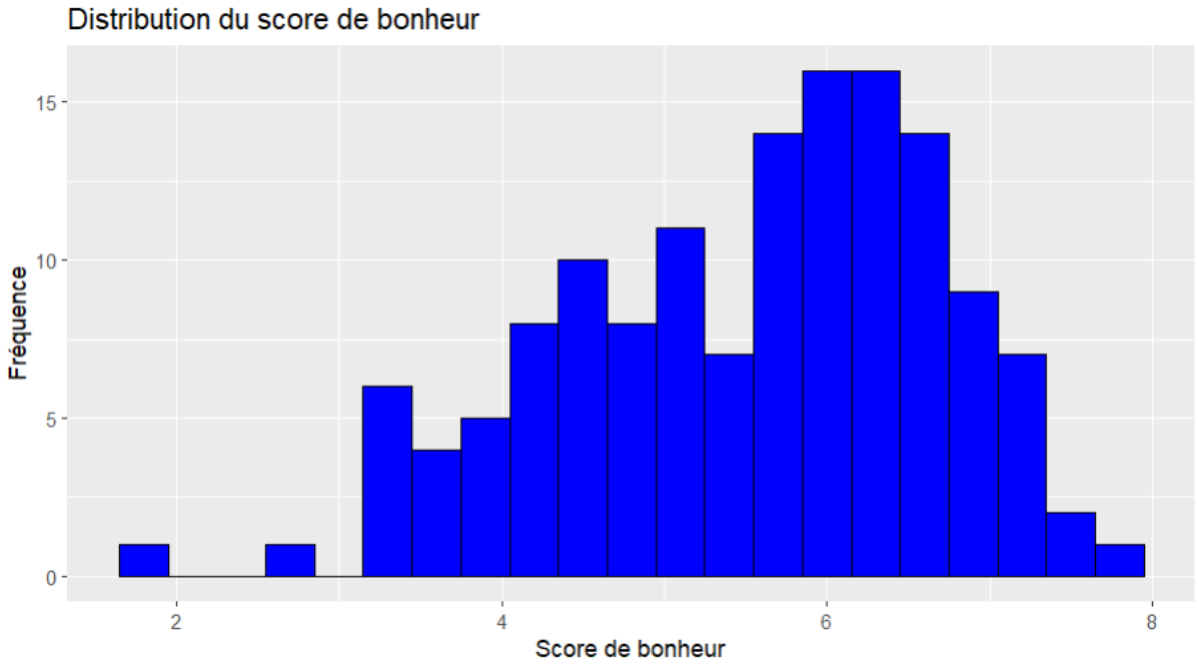
Code R :

```
ggplot(df_clean, aes(x = Score)) +
geom_histogram(binwidth = 0.3, fill = "skyblue", color =
"black") +
labs(title = "Distribution du score de bonheur", x = "Score de
bonheur", y
= "Fréquence")
```

Distribution du score de bonheur dans le monde (2024)

Analyse :

Ce graphique représente la répartition des scores de bonheur (appelés *Ladder score*) pour les différents pays selon le *World Happiness Report2024*. On observe que la majorité des pays ont un score compris entre 5 et 7, ce qui reflète une concentration autour d'un niveau de bonheur moyen à élever. Quelques pays présentent des scores très bas (autour de 2) ou très élevés (proches de 8).



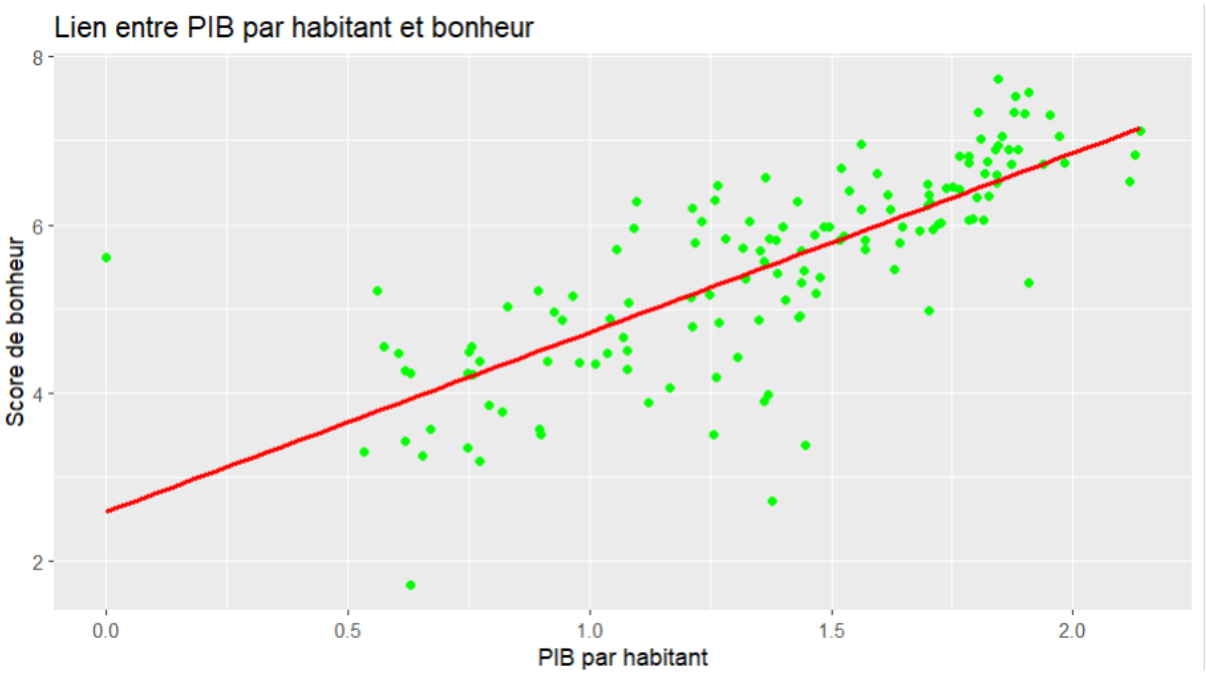
b) Nuage de points : PIB vs Score

Analyse :

Ce nuage de points montre la relation entre **le PIB par habitant (en log)** et **le score de bonheur** pour les différents pays.

On observe une **relation linéaire positive** : plus le PIB par habitant augmente, plus le niveau de bonheur tend à être élevé.

La droite rouge représente le modèle de **régression linéaire simple**, confirmant cette tendance globale, bien qu'une dispersion persiste entre les pays.



c) Boxplot : générosité

Code R :

```
df_clean$GroupeGénérosité <- ifelse(df_clean$Générosité >
  median(df_clean$Générosité, na.rm = TRUE), "Généreux", "Moins généreux")

ggplot(df_clean, aes(x = GroupeGénérosité, y = Score, fill = GroupeGénérosité)) +
  geom_boxplot() +
  labs(title = "Score de bonheur selon la générosité", x = "", y = "Score de bonheur") +
  theme_minimal()
```

Analyse : Peu de différence entre les pays généreux et moins généreux, confirmant que ce facteur influence peu le score global.

Score de bonheur selon la générosité des pays

Ce boxplot compare les scores de bonheur pour deux groupes : les pays plus généreux que la médiane et ceux moins généreux. Il permet de visualiser si la générosité perçue est un facteur associé au bien-être. On observe si les pays plus généreux ont tendance à être également plus heureux.

V. Conclusion

Quels sont les facteurs qui influencent le plus le bonheur d'un pays ?

Au terme de notre analyse basée sur les données du *World Happiness Report 2024*, plusieurs enseignements majeurs se dégagent quant aux déterminants du bonheur à l'échelle mondiale.

Les résultats mettent en évidence que **quatre variables principales** exercent une influence particulièrement forte et significative sur le score de bonheur :

- **Le PIB par habitant** : plus les ressources économiques moyennes d'un pays sont élevées, plus le niveau de bonheur déclaré tend à l'être également. Cet indicateur reflète l'accès aux biens, aux services et à la sécurité matérielle, des éléments qui contribuent au bien-être individuel.
- **Le soutien social** : la perception de pouvoir compter sur autrui en cas de besoin est un facteur central du bonheur. Il s'agit d'un levier profondément humain, qui montre que le lien social est aussi fondamental que les conditions économiques.
- **L'espérance de vie en bonne santé** : ce facteur traduit la qualité des systèmes de santé et le niveau général de bien-être physique. Un corps en bonne santé est logiquement un socle important du bien-être subjectif.
- **La liberté de faire ses propres choix** : le sentiment d'autonomie et de contrôle sur sa vie personnelle apparaît comme un facteur très significatif. Les pays où les citoyens se sentent libres de leurs décisions tendent à être plus heureux.

Ces quatre dimensions se distinguent aussi bien par leur **corrélation élevée avec le score de bonheur**, que par leur **poids significatif dans le modèle de régression multiple**, ce qui renforce la robustesse de ces conclusions.

À l'inverse, d'autres variables comme **la générosité** et la **perception de la corruption** se sont révélées **moins déterminantes**. Bien que ces aspects puissent jouer un rôle moral ou symbolique, leur impact statistique sur le niveau de bonheur semble plus marginal dans notre modèle. Cela ne signifie pas qu'ils sont inutiles ou sans valeur, mais qu'ils ne suffisent pas à eux seuls à prédire le bien-être global d'un pays.

VI. Annexe : Script R

```
#1. Installer les packages nécessaires


install.packages(c("readxl", "dplyr", "ggplot2", "corrplot",
"psych"))


#2. Charger les bibliothèques


library(readxl)

library(dplyr)

library(ggplot2)

library(corrplot)

library(psych)


# 3. Nettoyer l’environnement


rm(list = ls())


# 4. Importer les données traduites
setwd("C:/Users/Yacine/OneDrive/Bureau/SAE PROJET FICHIER")  #
adapter si besoin

df_raw <- read_excel("Data_WHR2024.xls")

# Vérifier structure

head(df_raw)

colnames(df_raw)


#5. Nettoyage (suppression lignes incomplètes)


df_clean <- df_raw[rowSums(is.na(df_raw)) <= 2, ]
```

```
# (Optionnel) sauvegarde

write.csv(df_clean, "data_WHR2024_clean_FR.csv", row.names = FALSE)


# 6. Statistiques descriptives


summary(df_clean)

describe(df_clean)



# Moyenne du bonheur

mean(df_clean$Score, na.rm = TRUE)


# 7. Corrélations entre variables

vars <- df_clean[, c("Score", "PIB/Habit", "Soutient Social",
"Éspérence de Vie",

                     "Liberté", "Générosité", "Corruption")]



cor_matrix <- cor(vars, use = "complete.obs")

print(cor_matrix)



corrplot(cor_matrix, method = "circle", type = "upper", tl.cex =
0.8)


# 8. Régression linéaire simple


modele_simple <- lm(Score ~ `PIB/Habit`, data = df_clean)

summary(modele_simple)
```

```
# 9. Régression multiple

modele_multiple <- lm(Score ~ `PIB/Habit` + `Soutient Social` +
`Éspérence de Vie` +
                        Liberté + Générosité + Corruption, data =
df_clean)
summary(modele_multiple)
# 10. Graphiques avec ggplot2
# Histogramme
ggplot(df_clean, aes(x = Score)) +
  geom_histogram(binwidth = 0.3, fill = "blue", color = "black") +
  labs(title = "Distribution du score de bonheur",
       x = "Score de bonheur",
       y = "Fréquence")
# Nuage de points + droite de régression
ggplot(df_clean, aes(x = `PIB/Habit`, y = Score)) +
  geom_point(color = "green") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Lien entre PIB par habitant et bonheur",
       x = "PIB par habitant",
       y = "Score de bonheur")
# Boxplot selon la générosité
df_clean$GroupeGénérosité <- ifelse(df_clean$Générosité >
                                      median(df_clean$Générosité,
na.rm = TRUE),
                                    "Généreux", "Moins généreux")
```

```
    ggplot(df_clean, aes(x = GroupeGénérosité, y = Score, fill =
                         GroupeGénérosité)) +

  geom_boxplot() +

  labs(title = "Score de bonheur selon la générosité",

       x = "Groupe de générosité",

       y = "Score de bonheur") +

  theme_minimal()

# 11. Visualisation des coefficients de la régression multiple

# On extrait tous les coefficients du modèle multiple
coefs <- summary(modele_multiple)$coefficients

# On les transforme en tableau (data frame) pour pouvoir les
manipuler
coef_df <- data.frame(
  Variable = rownames(coefs),
  Coefficient = coefs[, "Estimate"]        )

# On enlève l'intercept car ce n’est pas une variable explicative
coef_df <- coef_df[coef_df$Variable != "(Intercept)", ]

# On ajoute une colonne pour savoir si l'effet est positif ou
négatif
coef_df$Effet <- ifelse(coef_df$Coefficient > 0, "Positif",
"Négatif")

# On trace un graphique en barres pour voir l’effet de chaque
variable
library(ggplot2)

ggplot(coef_df, aes(x = reorder(Variable, Coefficient), y =
Coefficient, fill = Effet)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values = c("Positif" = "blue",
                               "Négatif" = "red")) +
  labs(title = "Effet des variables explicatives sur le score de
bonheur",
       x = "Variable explicative",
       y = "Coefficient de régression") +
  theme_minimal()
```