World Happiness Report Analysis

Introduction

The dataset used in this analysis comes from the World Happiness Report, an annual survey that assesses global well-being by collecting data on life satisfaction among other characteristics across different countries and years. Individuals rate their life satisfaction on a 0 to 10 scale. The dataset also includes a range of socioeconomic and psychological indicators, such as Log GDP per capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, Perceptions of Corruption, and Emotions. Detailed explanation of each variable can be found below:

- **Country Name**: The name of the country

- **Year**: The year of the observation (2005-2022)

- **Life Ladder**: This is the main happiness score (often called the Cantril ladder score), where people rate their life satisfaction on a scale typically from 0 to 10

- **Log GDP per Capita**: The natural logarithm of GDP per capita, which measures economic wealth

- **Social Support**: A measure of whether people have someone to count on in times of trouble. Values range from 0 to 1

- **Healthy Life Expectancy at Birth**: Average number of years a newborn can expect to live in good health

- **Freedom to Make Life Choices:** A measure of people's perception of their freedom to make their own life decisions. Values range from 0 to 1

- **Generosity**: A measure of charitable donation and helping behavior

- **Perceptions of Corruption**: How people perceive corruption levels in their government and businesses. Values range from 0 to 1

- **Positive Affect**: A measure of positive emotions like joy, laughter, and enjoyment. Values range from 0 to 1

- **Negative Affect**: A measure of negative emotions like worry, sadness, and anger. Values range from 0 to 1

By analyzing trends in Life Ladder scores over time, we aim to understand how economic, social, and psychological factors contribute to subjective well-being on individuals in a country. This dataset provides a broad cross-country perspective, allowing for comparisons between nations and different income groups over time. Therefore, we can use it to identify global trends in happiness and highlight the most important factors that influence variation in happiness over time.

Question of Interest

This project seeks to understand **which factors have the greatest impact on Life Ladder** (happiness) scores across countries and **how global crises influence overall well-being**.

Happiness is influenced by a combination of economic, social, and health-related factors, but it is hard to pinpoint the relative importance of each one. We can understand how real-world policies affect happiness by identifying the strongest predictors of Life Ladder scores. We can also make predictions based on a theoretical model that quantifies each predictor for future purposes. Therefore, a satisfactory answer may highlight the variables (e.g. GDP per Capita, Social Support, Corruption Perceptions) that are most helpful and hurtful when assessing Life Ladder scores.

This project will focus on the 2008 Global Financial Crisis and the 2020 COVID-19 pandemic as examples of crises, attempting to reveal quantifiable differences in happiness globally. A satisfactory answer would involve analyzing trends in happiness scores in the years before and after 2008 and 2020. If significant decreases in Life Ladder scores or other variables coincide with crises, we can conclude that global shocks have a measurable impact on well-being or behaviors. Lastly, an answer to how different income groups or regions respond to crises is relevant. Some countries may have more volatile declines in happiness than others, depending on economic or social infrastructure.

Data Preparation

The World Happiness Report dataset, initially containing 2,200 observations, was imported and cleaned by removing any missing or NA values. This may introduce some bias because the data was not missing at random and it is likely that older observations of lower-income countries get disproportionately removed. However, this is necessary for regression and graphical modeling, as the missing values would otherwise cause errors. The dataset still consisted of around 2000 observations which was used for the rest of the analysis.

Exploring the Relationship Between Key Factors and Happiness

Correlation Matrix

The first step I took was to explore more general relationships between Life Ladder and other variables in the dataset. This was done through a correlation matrix.

Life Ladder Correlation Matrix
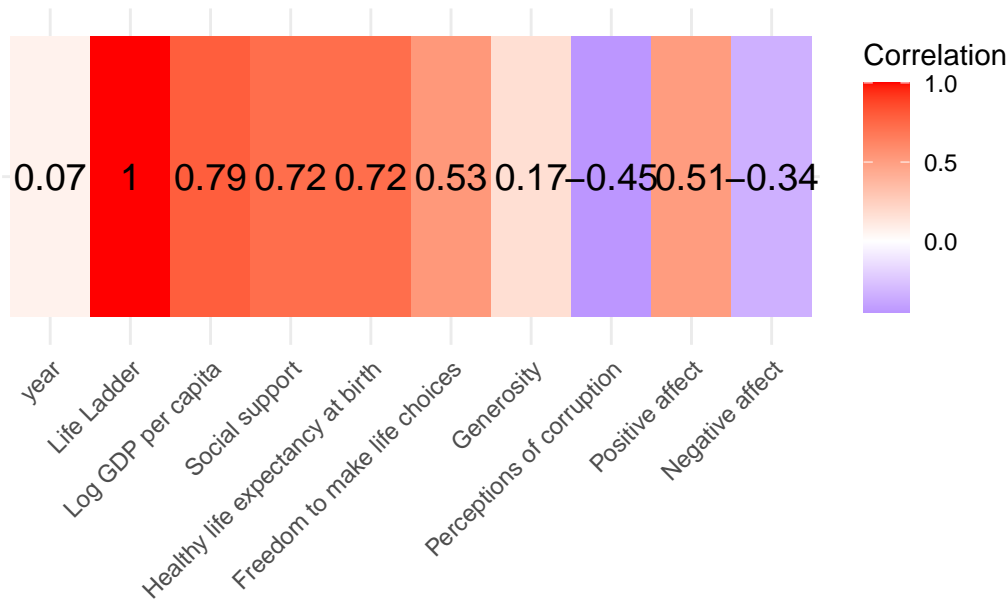
This correlation visualization helped identify which variables are most strongly correlated with Life Ladder scores.

- Life Ladder has a **strong positive correlation with Log GDP per capita** (0.79), suggesting that as GDP increases, quality of life and happiness also improves

- **Social support** (0.72) and **Healthy life expectancy at birth** (0.72) also show **strong positive correlations** with Life Ladder, indicating that higher levels of social support and better health outcomes are associated with higher life satisfaction

- Freedom to make life choices (0.53) **positively correlates** with Life Ladder

- Perceptions of corruption exhibits a **moderate negative correlation** (-0.45), revealing that as perceived corruption increases, life satisfaction decreases

- Generosity (0.17) and Positive affect (0.51) have **weaker positive correlations**, while Negative affect (-0.34) shows a **moderate negative correlation** with Life Ladder

This plot gave me a foundation for which to do more specific research and an idea for which variables I should include when doing regression.

Life Ladder and Quartiled Data

To understand how different socioeconomic variables relate to happiness, I grouped countries into quartiles based on their Log GDP per Capita, Social Support, Freedom to Make Life Choices, and Perceptions of Corruption.

I created line graphs that tracked the average Life Ladder score over time for each quartile. This allowed me to observe how different income levels or social conditions influence happiness trends and whether disparities exist between groups.

Average Life Ladder by Characteristics
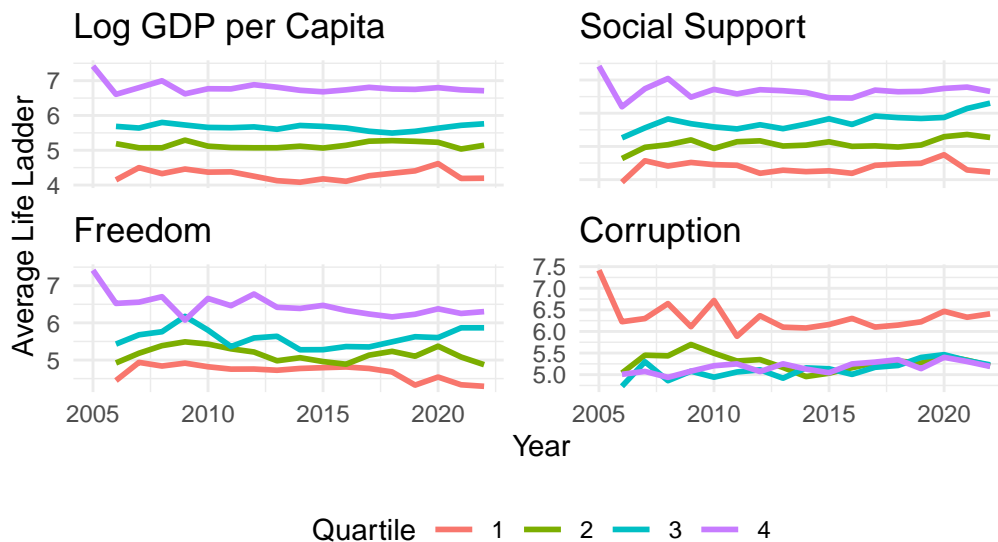
From this group of plots we can observe the effect that Log GDP per Capita, Social Support, Freedom to Make Choices, and Perceptions of Corruption have on Life Ladder.

Log GDP per Capita: Being in the top or bottom quartile of income has a large effect on happiness, as evidenced by the gap between the fourth and first quartiles from the third and second quartiles respectively.

Social Support: This graph exhibits similar attributes as the income graph, but the gaps in happiness between quartiles are smaller. Still, greater access to social support has a positive correlation with happiness.

Freedom to Make Choices: The difference in happiness between quartiles in this graph is quite small, but volatile over time. Feelings of freedom were less correlated with Life Ladder in the correlation matrix seen earlier, and this graph highlights how it certainly has an effect on Life Ladder, but it is not as prominent as income or social support.

Perception of Corruption: Countries with the lowest perceptions of corruption have the highest happiness. Interestingly, the next three quartiles have almost no difference in happiness,

maybe revealing some sort of non-linear slippery slope effect. Once someone believes there is even a little corruption in a country, their happiness significantly decreases as a result.

Overall, by grouping countries by different quartiles, it becomes clear that **high income, social support, and freedom are important to higher Life Ladder ratings**. Extremely **low perceptions of corruption** are important as well, as even average amounts of corruption perception have a drastic negative effect on happiness.

Multiple Linear Regression

Lastly, I performed a multiple linear regression analysis, using Life Ladder as the response variable and the previous socioeconomic indicators as explanatory variables. The regression equation is as follows.

<div align="center">

Life Ladder Multiple Regression

</div>

Predictor	Estimate	Std. Error	t Value	p-Value
(Intercept)	23.995	5.973	4.017	6.12e-05
Year	-0.013	0.003	-4.339	1.51e-05
Log GDP per capita	0.341	0.023	14.930	< 2e-16
Social Support	2.518	0.154	16.352	< 2e-16
Healthy Life Expectancy	0.028	0.003	8.376	< 2e-16
Freedom to Make Life Choices	1.492	0.117	12.789	< 2e-16
Perceptions of Corruption	-0.814	0.083	-9.869	< 2e-16

All of the following interpretations of variables are being done under the assumption that all over variables are being held constant except for the one being highlighted. Each one of these variables have a p-value less than 0.05, meaning they are significant at the 95% level.

Intercept: This intercept of 23.995 should not be interpreted because it represents the happiness score when all other variables are zero. Life Ladder scores are on a scale from 0-10, so a value of 23 is not possible. Another reason the intercept should not be interpreted is because it is impossible for all explanatory variables to be zero in this model.

Year: The coefficient of -0.01 means that every year, happiness levels tend to decrease slightly. This may seem insignificant, but the significant p-value shows that we cannot ignore this trend.

Log GDP per Capita: The coefficient of 0.34 shows how there is a strong positive correlation between GDP per capita and Life Ladder. Since this is log transformed, a 1% increase in GDP per capita represents a 0.34 increase in happiness.

Social Support. The coefficient of 2.518 is the largest in the model, but the domain for social support in our dataset is from 0 to 1. This means that a 0.01 or 1% increase in social support represents a 0.025 increase in happiness, slightly less than GDP per capita. In extreme cases, going from no social support to complete social support predicts a 2.518 increase in life ladder.

Healthy Life Expectancy: The coefficient of 0.028 is small, but can directly be interpreted. A 1 year increase in healthy life expectancy is predicted to increase happiness by 0.028. This is a greater effect than social support. In our correlation matrix, life expectancy and social support had the same correlation, but this model reveals that life expectancy is a slightly stronger explanatory variable.

Freedom to Make Choices: Due to this variable being ranged from 0 to 1, a 1% increase represents a 0.015 increase in happiness. This is still positively correlated, but much weaker than our previous explanatory variables.

Perceptions of Corruption: This is the main negative correlated variable with Life Ladder. Similar to before, the 0 to 1 scale of this variable means that a 1% increase in perceptions of corruption has a -0.008 effect on predicted Life Ladder. This linear explanation may not be fully accurate, since we learned earlier that there exists a steep drop-off on how corruption perceptions affect happiness. However, we can still understand the relative correlation of this variable through the regression model.

Diagnostic Checks for Multiple Linear Regression Model

The assumptions for regression analysis regarding residuals must be checked for the model to be significant and valid to draw conclusions from (diagnostic plots in appendix #1). The first three diagnostic plots of the multiple linear regression model indicate that **the residuals follow the necessary assumptions of normality, linearity, and homoscedasticity** (constant variance). The Residuals vs. Fitted Plot shows no clear pattern in the residuals, suggesting that the relationship is likely linear. The Q-Q Plot demonstrates that the residuals are normally distributed, with the points closely following the a straight line. The Scale-Location Plot indicates constant variance across the fitted values, supporting the assumption of homoscedasticity (constant variance).

However, the Residuals vs. Leverage Plot identifies three influential points. 710, 711, and 713 are labelled as influential points, which means that the inclusion of these points has a significant effect on our model and must be looked into. These three points represent data from Haiti when they experienced a 7.1 magnitude earthquake. These points significantly lowered the country's healthy life expectancy at birth for those years.

To find how much the outlier points truly influenced the model, I ran an adjusted regression model that excluded these three points. The results of this adjusted model (presented in the appendix #2) show few differences. The intercept and healthy life expectancy at birth

coefficients are slightly higher in the adjusted model, indicating that the outliers skewed the results in the original model. However, all other explanatory variables are constant, and the R-squared values in the adjusted model is within 1% of the original model.

While these points are outliers, I chose to retain them in the analysis and main regression model because they reflect an important part of Haiti's history and the long-term impact of the earthquake on its population, making their inclusion essential for identifying global trends.

Our regression model passes all diagnostic checks and assumptions, so we are able to draw conclusions and use it to predict life ladder based on the explanatory variables.

The original regression model made it clear which variables have the greatest **absolute** impact on Life Ladder scores.

1. Log GDP per Capita (0.34)
2. Healthy Life Expectancy (0.028)
3. Social Support (0.025)
4. Freedom to Make Choices (0.015)
5. Perceptions of Corruption (0.008)
6. Year (0.01)

Global Crises and Happiness Levels

This section examines how happiness levels have evolved over time, particularly during major global crises such as the 2008 financial crisis and the COVID-19 pandemic in 2020. I wanted to explore how people respond to global crises and if there are any drastic socioeconomic changes resulting from the shocks in the economy.

Global Average Life Ladder Before and After Crisis Years

The first thing I checked was how life ladder values change the year before and after a crisis. I hypothesized that there would be a decrease in average life ladder values after a global crisis. However, the results (seen in the appendix #3) did not fully agree with this belief.

- **Quartile 1 (Lowest Income)**: Life ladder decreased slightly by 0.2 from 2007 to 2008 and rose back up 0.15 in 2009. The opposite effect occurred in 2020, as life ladder increased by 0.2 but dropped by over 0.4 in 2021. This was initially interesting and required more research to understand why life ladder shot up in one crisis but down in another.

- **Quartile 2**: Life ladder stayed exactly constant from 2007 to 2008 and increased by 0.2 in 2009. A similar constant effect can be observed in 2019 to 2020. However, a decrease of 0.2 followed in 2021 rather than an increase.

- **Quartile 3**: Life ladder increased from 2007 to 2008 by 0.15 and fell by 0.07 in 2009. From 2019 to 2020, life ladder increased by 0.1 and continued to increase in 2021 by another 0.1.

- **Quartile 4 (Highest Income)**: Life ladder increased by 0.2 from 2007 to 2008, but then fell by around 0.4 in 2009. From 2019 to 2021, life ladder for the top quartile stayed constant, something that was unexpected and also required more research.

It seems that the changes in happiness are not exponentially correlated to lower income. It is also unknown whether these changes in happiness are due to the crisis, outside factors not included in this dataset, or random variation in happiness levels across years. I wanted to analyze more than life ladder and look at trends much further than one year before and after the crisis. It is hard to tell whether average life ladder values decreased as a result of global crises simply from this table. By looking at many more variables across time, we can make a more accurate prediction to how different income brackets respond to crises.

Global Crises and Socioeconomic Factors

I chose to analyze Life Ladder, Generosity, Positive Emotions, and Negative Emotions to achieve a more sophisticated result. My hypothesis was that for higher income brackets, crises would slightly decrease Life Ladder and Positive Emotions, while increasing Generosity and Negative Emotions. For lower income brackets, crises would greatly decrease Life Ladder, Generosity and Positive Emotions, while increasing Negative Emotions.

I grouped countries by their income quartile and plotted average happiness, generosity, positive, and negative emotions over time. The two dashed black lines represent 2008 and 2020 respectively.

Characteristics by Income Quartile (2006–2022)
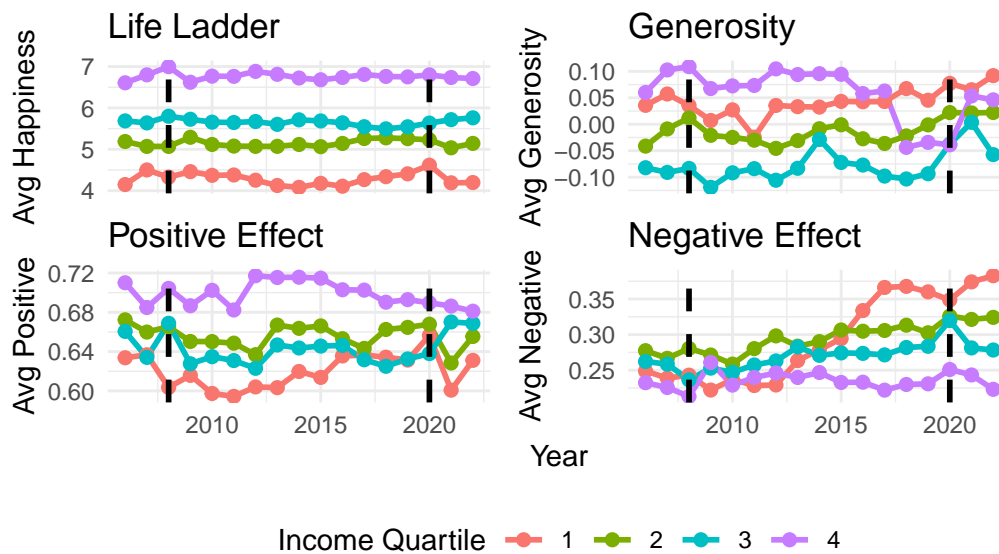
Life Ladder:

- In 2008, average life ladder increased in the top two quartiles, but slightly decreased in the bottom two quartiles. The most significant changes occured in the top and bottom quartiles, as the extreme wealth (or lack of wealth) likely were the reasons.

- In 2020, the top three quartiles were relatively unchanged, possibly showing a better adaptation to the COVID-19 pandemic after the 2008 financial crisis. Surprisingly, the bottom quartile had an increase in life ladder, possibly a result of the financial aid being given out having a significant effect on the lower income levels.

Generosity:

- In 2008, each quartile had a minor decrease in generosity after the financial crisis. This shows how people might have been more focused on their own finances rather than helping out others.

- In 2020, generosity either stayed stable or slightly increased across all quartiles. This could reflect an upsurge in community support and charitable giving in response to the pandemic, once again a sign of learning from 2008.

Positive Emotions:

- In 2008, each quartile showed a small decrease in positive emotions either during the crisis or shortly after.

- However, in 2020, the top two quartiles were seemingly unaffected while the bottom two quartiles had a large decrease in positive emotions. This may be due to wealthier people being more optimistic about the future during times of crisis.

Negative Emotions:

- In 2008, the top two quartiles had increases in negative emotions, with the top quartiles being the largest by far. The bottom two quartiles had more stable averages.

- In 2020, negative emotions spiked during the pandemic, but quickly decreased the following year for all income quartiles except for the lowest. The bottom quartile had the opposite effect, with negativity decreasing during 2020, but spiking upwards the following year. Once again, this may be due to the stimulus and outside help that poorer countries were receiving throughout the pandemic.

Rationale and Thoughts

In 2008, the financial crisis led to noticeable negative effects on life satisfaction, generosity, positive emotions, and an increase in negative emotions, especially for the wealthiest and the poorest countries. However, the situation in 2020 was quite different, possibly due to the rapid government responses and support systems put in place in many countries. In the wake of the COVID-19 pandemic, the wealthier quartiles were largely unaffected in terms of Life Ladder, Generosity, and Positive Emotions. This could reflect the ability of wealthier populations to adapt more easily due to their access to resources. The bottom quartile experienced an unexpected increase in Life Ladder in 2020, likely due to the financial assistance provided by governments to support those hit hardest by the pandemic.

These findings were a stark contrast to my original hypotheses. My original proposal was that global crises will have a noticeable negative effect on life ladder, but I had failed to understand the role of government in countries going through crises. When "times are good", governments will attempt to take measures that provide some form of insurance in the case of extreme negative shocks occurring. When these come to fruition, governments will use the resources they have gathered to try and smooth out these socioeconomic factors to keep the economy and society moving. This data shows that governments are still learning about how to deal with financial crises, and they are continuing to do a better job at managing both economic and social variables to prevent them from crashing. The COVID-19 pandemic was one of the largest global economic crashes, yet the life ladder and other measurements show only slight negative effects.

Summary of Findings and Discussion

This project aimed to answer two key questions:

- Which factors have the greatest impact on Life Ladder (happiness) scores across countries?

- How do global crises influence overall well-being?

The analysis focused on understanding key drivers of happiness and how the 2008 Financial Crisis and the COVID-19 pandemic affected happiness and other characteristics around the world.

To determine the most influential factors on Life Ladder scores, we first created a correlation matrix, which revealed that Log GDP per Capita, Social Support, and Healthy Life Expectancy at Birth had the strongest positive relationships with happiness. Perceptions of Corruption showed a moderate negative correlation, suggesting that trust in government is also important for general happiness. Generosity and Positive Emotions had weaker correlations but still contributed to overall well-being, while negative emotions had weaker negative correlation. Ultimately, **economic stability, strong social networks, and strong medical systems that ensure long healthy lives are the most significant contributors** to happiness across countries. These findings match my original hypothesis, as I expected richer, more connected countries with long lives to report the greatest happiness scores.

Next, to examine the impact of global crises on happiness, we grouped countries into quartiles based on their income and tracked changes in key happiness indicators before and after the 2008 financial crisis and 2020 COVID-19 pandemic. The results showed that **happiness scores in the top and bottom quartiles were the most affected by economic downturns**, while middle-income groups remained more stable. Wealthier countries could use their established wealth during tight economic times, essentially preventing a majority of negative effects experienced by poorer countries. After the 2008 crisis, generosity and positive emotions declined globally as a result of the stress and pessimism experienced by the economic crash. However, during the COVID-19 pandemic, the bottom quartile unexpectedly saw an increase in Life Ladder scores, possibly due to financial aid programs that temporarily boosted well-being. The other three quartiles saw little change in happiness reports. Interestingly, generosity increased in all quartiles after 2020. The pandemic was a time of isolation for many people, so it is feasible that once the economy reopened people felt like helping one another recover. Negative emotions rose during 2020, but decreased the following year with an increase in positive emotions as well.

These findings did not exactly match my hypothesis that crises would decrease happiness scores and other positive characteristics. I failed to assess the role of government and how they mitigate many effects of crises globally. This was an eye-opening insight that increased my appreciation for the importance of government, as people only seem to mention the failures of the government rather than all of the successes.

Overall, these findings highlight that **global crises do impact well-being, but government responses and economic safety nets play a crucial role in mitigating negative effects**. Unlike the 2008 crisis, where financial hardship led to broad declines in happiness,

the rapid government intervention during COVID-19 helped prevent a severe drop in Life Ladder scores. This suggests that proactive policies, such as financial relief programs, can help maintain stability during economic downturns.

Limitations and Recommendations

While this analysis provides valuable insights, there are several limitations to consider. The removal of missing values may have disproportionately excluded lower-income countries from the analysis, potentially biasing results. The dataset also does not account for differences in data collection methods across countries, which may introduce inconsistencies in Life Ladder scores. Additionally, the Cantril ladder score is a subjective ranking, which means that objective analysis of happiness is not possible. The methods used to answer questions in this project were also very general, and this analysis fails to look at specific trends between individual countries over time. The averaging of factors removes the nuance from many observations, which should be considered when seeking to answer similar questions.

One recommendation I have for others attempting to answer similar questions would be to analyze additional variables such as education levels, unemployment rates, mental health metrics, or crime rates. These metrics could provide a more comprehensive understanding of what influences happiness beyond the socioeconomic factors in this project. Exploring the interaction effects between variables (e.g., how corruption perception changes the impact of GDP on happiness) could uncover more sophisticated or realistic results. Lastly, using other methods of analysis such as weightings for each variable or grouping by metrics other than income quartile may uncover different results for deeper analysis or conclusions.

Appendix

1. MLR Model Diagnostics
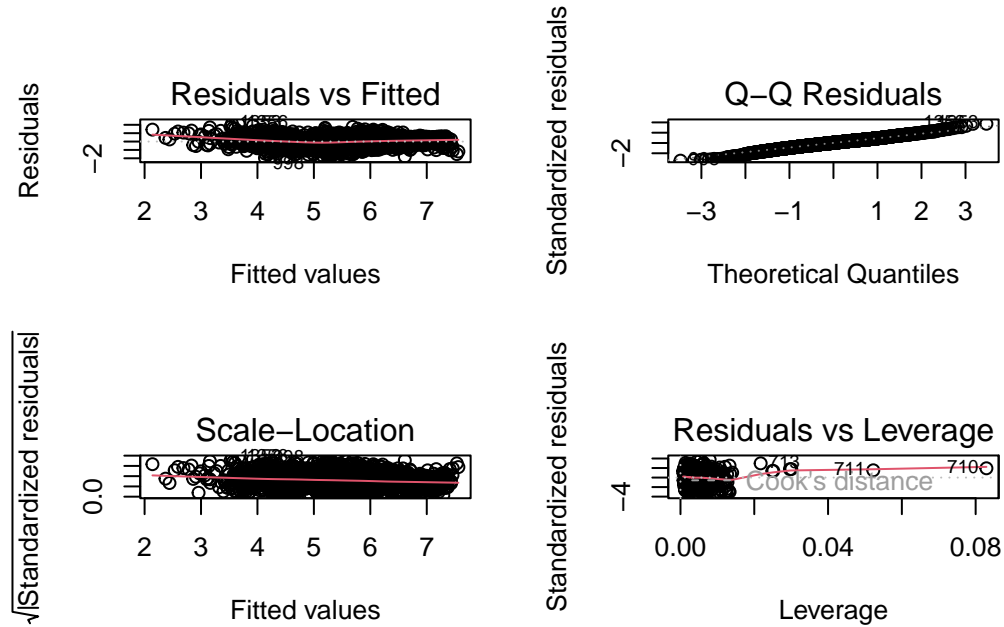
2. MLR Model with Removed Outliers

Life Ladder Multiple Regression

Predictor	Estimate	Std. Error	t Value	p-Value
(Intercept)	24.730	5.956	4.152	3.44e-05
Year	-0.013	0.003	-4.494	7.39e-06
Log GDP per capita	0.318	0.024	13.478	< 2e-16
Social Support	2.515	0.154	16.383	< 2e-16
Healthy Life Expectancy at Birth	0.033	0.004	9.119	< 2e-16
Freedom to Make Life Choices	1.514	0.116	13.006	< 2e-16
Perceptions of Corruption	-0.802	0.082	-9.741	< 2e-16

3. Average Life Ladder by Income Quartile

Average Life Ladder by Income Quartile

Q1 (Lowest)	Q2	Q3	Q4 (Highest)
2007			
4.50	5.07	5.64	6.80
2008			
4.33	5.07	5.80	7.00
2009			
4.46	5.29	5.73	6.62
2019			
4.41	5.26	5.54	6.75
2020			
4.62	5.23	5.63	6.80
2021			
4.19	5.04	5.72	6.74