Analysis of Global Happiness Factors

Ethan Tsao

Table of contents

1 Abstract

This report analyzes global well-being trends using the World Happiness Report dataset, starting with how GDP and other economic indicators impact happiness across countries. The study investigates how these economic factors, including GDP per capita, social support, and healthy life expectancy, interact with happiness levels in different regions. Following this, predictive models such as random forests and linear regression are used to estimate the top 10 happiest countries in 2025 based on life ladder scores. The analysis shifts focus to life expectancy, providing a comparative study between countries at different developmental stages, with particular attention to the United States and Mexico, highlighting how economic conditions influence health outcomes. Lastly, we will explore regional differences in emotional well-being, examining positive and negative affect levels across Asia, Europe, and North America from 2005 to 2020, using ANOVA tests to identify significant regional disparities. Through various visualizations and hypothesis testing, the report emphasizes the role of economic stability, social policies, and cultural influences in shaping overall happiness worldwide.

1.1 Question of Interest

Happiness is often thought to be influenced by a combination of economic, social, and health-related factors, but the extent to which each of these contributes to overall well-being varies across countries. We will explore how key factors such as GDP per capita, social support, and healthy life expectancy influence happiness levels globally. By analyzing data from the World Happiness Report, we seek to determine which factors have the strongest correlation with happiness, whether these relationships vary by region, and how they have evolved over time. Understanding these patterns can provide insight into the most effective ways to improve well-being on a national and global scale.

1.2 Sub-Questions

1. How do GDP and other factors related to happiness interact across different countries?
2. Can we predict what the top 10 happiest countries will be in 2025 based off of life ladder score?
3. How does life expectancy differ between countries in different developmental stages?
4. What specific factors contribute to the observed differences in positive and negative affect levels between Asia, Europe, and North America?

1.3 Variable Descriptions

- **country_name**: The name of the country that the data was collected in.
- **year**: The year the data was collected in (2008-2023).

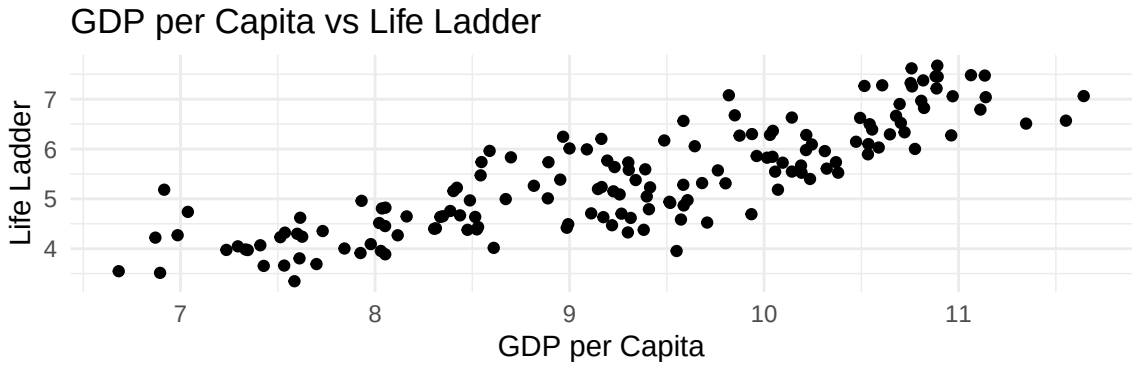
- **life_ladder**: The main happiness score, measured on a scale from 0 to 10 based on how people within a country perceive and evaluate their well-being/quality of life.
- **gdp_per_capita**: A measure of economic wealth that reflects the average income level of individuals in a country adjusted based on population size.
- **social_support**: A measure of whether individuals feel they have someone to rely on in times of need on a scale from 0 to 1.
- **healthy_life_expectancy**: A measure that indicates the overall health and well-being of the country's population.
- **freedom_of_choice**: A measure of individual autonomy and personal freedoms in making life choices
- **generosity**: A measure of charitable donations that indicates helping behavior on average nationally.
- **perceptions_of_corruption**: How people perceive corruption levels in their government and business on a scale from 0 to 1.
- **positive_affect**: The average level of positive emotions in a country, based on survey responses about laughter, enjoyment, and engaging activities.
- **negative_affect**: The average level of negative emotions in a country, based on survey responses about worry, sadness, and anger.

2 Country GDP vs Happiness Factors

We start by cleaning the 2023 World Happiness Report dataset using the clean_names() function from the Janitor library. This standardizes column names by converting them to lowercase, replacing spaces and special characters with underscores, and ensuring they are syntactically valid. Additionally, we rename certain columns for clarity, making them easier to interpret in visualizations.

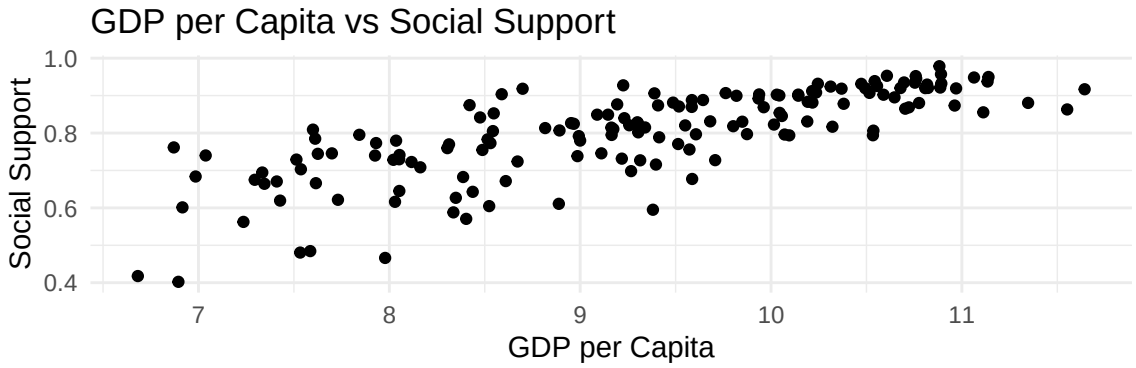
Gross Domestic Product (GDP) reflects the size and performance of a country's economy, serving as a key indicator of overall wealth. Since economic conditions influence many aspects of well-being, we explore how GDP relates to key happiness factors such as life ladder, social support, healthy life expectancy, generosity, and perceptions of corruption. We will aggregate the data by country to analyze the connection between GDP and happiness factors.

2.1 GDP v. Life Ladder



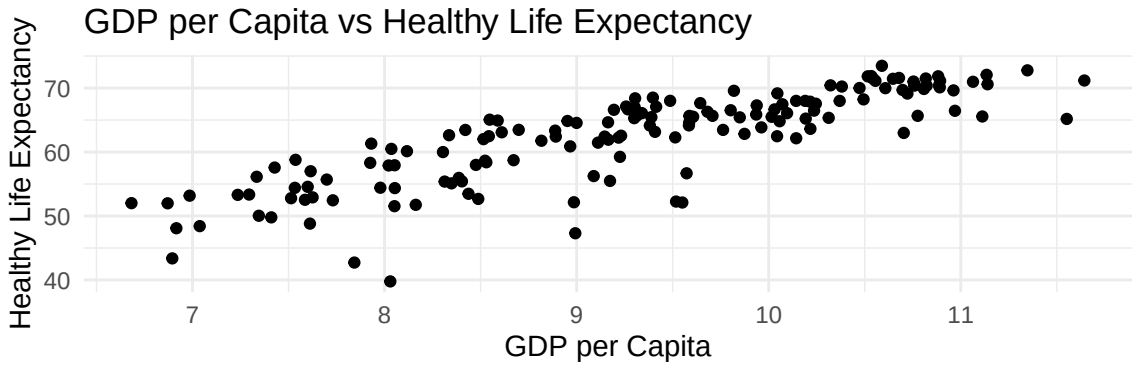
From the graph above, we observe a clear positive correlation between GDP per Capita and Life Ladder score. As GDP per Capita increases, so does the Life Ladder score, which is consistent with the idea that higher economic wealth often leads to an improved sense of well-being and quality of life. This relationship suggests that people in wealthier countries tend to rate their lives more positively, likely due to better living conditions, access to healthcare, and overall financial security.

2.2 GDP v. Social Support



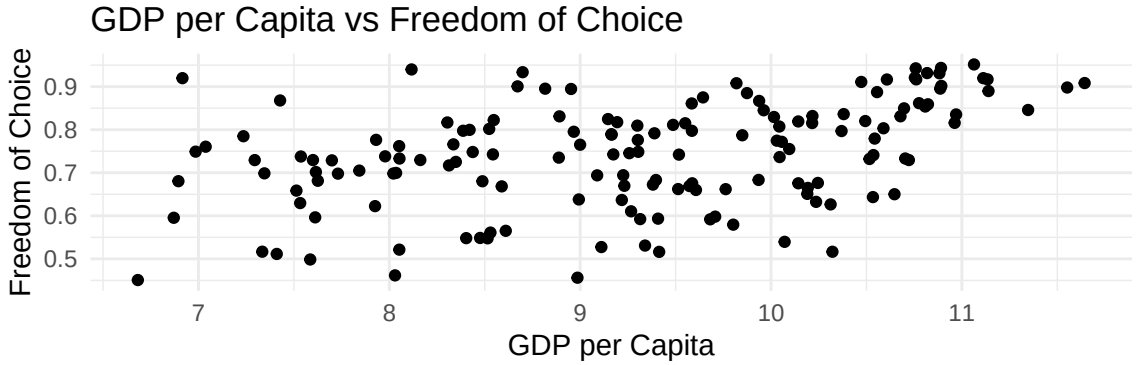
There appears to be a positive relationship between GDP per Capita and the Social Support score, with higher GDP generally corresponding to higher levels of social support. Social Support measures the extent to which people can rely on others, and it makes sense that increased financial stability would contribute to greater comfort and the ability to build social networks. When comparing this graph to the one depicting GDP versus Life Ladder, we observe that the correlation here is slightly weaker. The increase in Social Support is more pronounced at the lower GDP levels, but it starts to taper off as GDP per Capita rises, suggesting diminishing returns at higher income levels.

2.3 GDP v. Healthy Life Expectancy



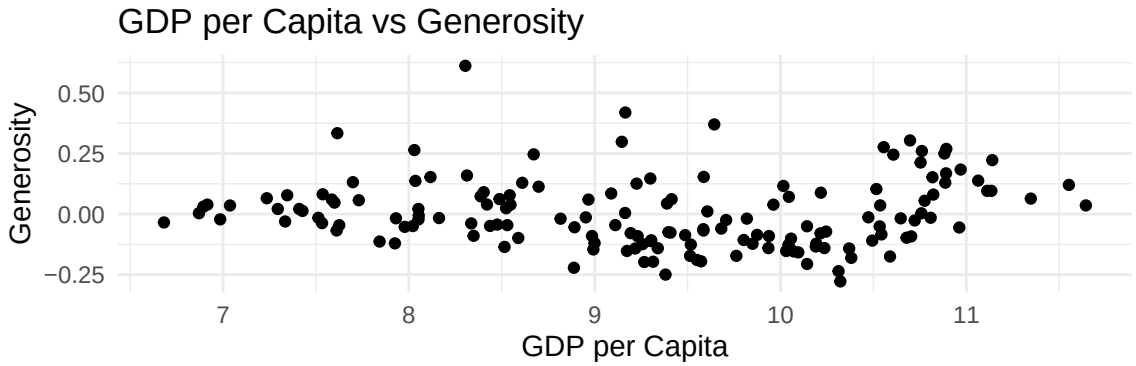
The graph clearly shows a positive correlation between GDP per Capita and Healthy Life Expectancy. As GDP per Capita increases, so does Healthy Life Expectancy. This likely reflects the fact that higher income levels contribute to improved living conditions, better access to healthcare, and healthier lifestyle choices, such as improved diet, sleep, and overall well-being. These factors combined tend to enhance life expectancy and health outcomes.

2.4 GDP v. Freedom of Choice



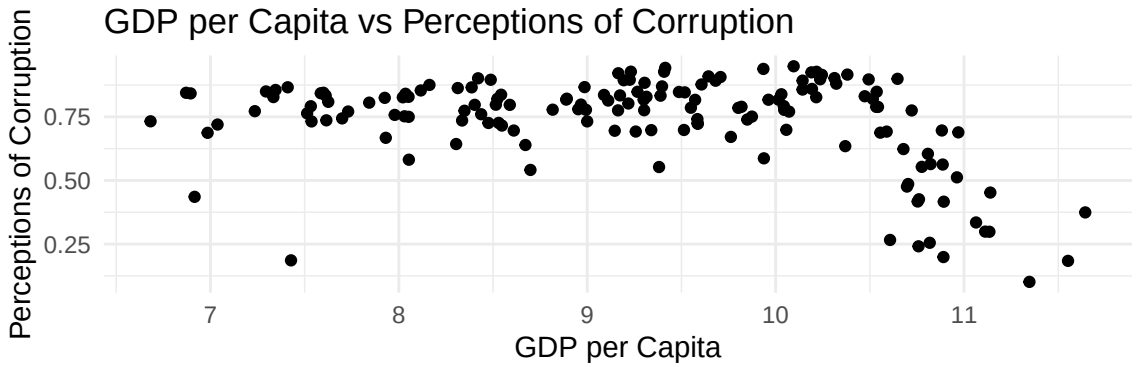
The scatterplot above shows no clear relationship between GDP per Capita and Freedom of Choice. While one might expect that higher "financial freedom" would correlate with greater overall freedom, this does not appear to be the case. It suggests that Freedom of Choice may be more strongly influenced by factors such as the political climate, government policies, or legal systems in different regions, rather than solely by economic wealth.

2.5 GDP v. Generosity



The scatterplot suggests that as GDP per Capita increases, generosity appears to slightly decrease. However, the relationship between the two seems weak, indicating little to no clear correlation. This could be attributed to the fact that national average charitable donations may not be heavily influenced by GDP per Capita alone, as generosity is likely shaped by a variety of other factors, such as cultural norms, social policies, and individual values.

2.6 GDP v. Perceptions of Corruption



The scatterplot indicates that as GDP per Capita increases, perceptions of corruption tend to decrease, particularly towards the higher end of the graph. In the lower and middle ranges of GDP, there appears to be little to no clear relationship between the two variables. However, once GDP reaches around 10, there is a sharp decline in perceptions of corruption. This suggests that higher GDP per Capita might be linked to improved living standards, and as people's basic needs are met, their focus may shift more toward the transparency and integrity of their government.

3 Statistical Modeling and Top 10 Life Ladder Scores Prediction

3.1 Linear Regression

Our EDA set the stage for analyzing how various factors interacted with GDP per Capita, but we will now turn our focus from visualization to statistical modeling. In this analysis, we will apply linear regression to explore the relationship between various factors and the Life Ladder score, which reflects the well-being of countries. Specifically, we aim to examine how factors such as GDP per capita, social support, healthy life expectancy, freedom of choice, generosity, and perceptions of corruption influence a country's Life Ladder score. Linear regression is a suitable method for this task because it allows us to quantify the strength and direction of these relationships, enabling a clearer understanding of which factors most significantly impact life satisfaction across different nations. Through this analysis, we hope to gain insights into how economic and social variables contribute to overall national well-being.

3.2 GDP per Capita Statistics

Table 1: Linear Regression Model Results

	Estimate	Std. Error	t Value	Pr(>|t|)
(Intercept)	2.310	0.557	4.145	0.000
life_ladder	0.452	0.088	5.113	0.000
social_support	2.225	0.603	3.691	0.000
healthy_life_expectancy	0.065	0.010	6.685	0.000
freedom_of_choice	-1.075	0.486	-2.213	0.028
generosity	-0.758	0.323	-2.344	0.020
perceptions_of_corruption	-0.675	0.293	-2.307	0.022

Our first linear regression model provides significant insights into how various factors influence GDP per capita as shown in our previous scatterplots. The most significant predictor is Healthy Life Expectancy, with a high t-value of 6.685 and an extremely small p-value (4.34e-10), suggesting that it has a strong positive effect on GDP per capita. Life Ladder also has a strong positive relationship with GDP per capita, with a t-value of 5.113 and a p-value of 9.60e-07, further emphasizing the importance of well-being measures in economic outcomes.

Other factors like Social Support and Freedom of Choice are also significant, with Social Support showing a positive relationship with GDP per capita (t-value of 3.691), and Freedom of Choice showing a negative relationship (t-value of -2.213). The negative relationships with Generosity and Perceptions of Corruption suggest that higher generosity and lower perceptions of corruption might correlate with lower GDP per capita in this dataset, but these relationships are still statistically significant (with p-values below 0.05).

Overall, the model has a strong fit, with an R-squared value of 0.8207, indicating that 82% of the variation in GDP per capita can be explained by the included variables

3.3 Life Ladder Correlation

Table 2: Correlations with Life Ladder

Factor	Correlation with Life Ladder
gdp_per_capita	0.8406568
social_support	0.7868488
healthy_life_expectancy	0.7819035
freedom_of_choice	0.6220699
generosity	0.1765117
perceptions_of_corruption	-0.4600152

The correlations between the various factors and the Life Ladder score provide valuable insights into what influences life satisfaction. The strongest correlation is with GDP per capita (0.841), indicating that as GDP per capita increases, so does the perceived life satisfaction. This relationship suggests that wealthier countries, which likely offer better living conditions, resources, and healthcare, tend to have higher life satisfaction. Social support (0.787) also shows a strong positive correlation, emphasizing the importance of strong social networks and the availability of support from family and friends in enhancing life satisfaction. Similarly, healthy life expectancy (0.782) is another important factor, indicating that people in healthier societies generally report higher levels of well-being.

The correlation with freedom of choice (0.622) is positive but weaker. This suggests that while the ability to make personal choices plays a role in life satisfaction, it is not as significant as factors like economic wealth or social support. Generosity shows a weak positive correlation (0.177), indicating that while generosity may contribute to life satisfaction, its impact is relatively small compared to the other factors. On the other hand, perceptions of corruption (-0.460) have a moderate negative correlation with the Life Ladder score. This suggests that as corruption increases, life satisfaction tends to decrease, likely due to the negative effects of corruption on governance, inequality, and social trust.

In summary, the strongest predictors of life satisfaction are economic wealth, social support, and health, with freedom of choice also playing a role, albeit a smaller one. While generosity has a minor influence, the presence of corruption appears to have a significant negative effect on life satisfaction, highlighting the importance of good governance and transparency for society's well-being.

3.4 Random Forest

Table 3: Random Forest Cross-Validation Results

mtry	RMSE	Rsquared	MAE
2	0.4576981	0.8460542	0.3547562
4	0.4660689	0.8367606	0.3555247
6	0.4719657	0.8322445	0.3619128

Table 4: Random Forest Model Evaluation Metrics

Metric	Value
RMSE	0.3731831
R-squared	0.8570122
MAE	0.3109879

From the Random Forest Cross-Validation Results, we can see the model's performance across different values of mtry. The lowest RMSE is observed when mtry = 2, with a value of 0.4577. Additionally, the highest R-squared value (0.8461) is also observed at mtry = 2. As mtry increases to 4 and 6, both RMSE and R-squared values worsen, suggesting that increasing the number of predictors doesn't necessarily improve the model's accuracy in this case. This indicates that the optimal number of predictors (mtry = 2) is well-suited for our dataset.

From the Random Forest Model Evaluation Metrics, the model shows strong performance, with an RMSE of 0.3732, indicating relatively low average prediction error. The R-squared value of 0.8570 is very good for our particular regression modeling. The MAE of 0.3110 indicates that the model's predictions, on average, are off by approximately 0.3110 units, further reinforcing the model's overall accuracy.

In conclusion, the Random Forest model appears to perform well on this dataset, as evidenced by low RMSE, high R-squared, and a relatively low MAE. The optimal value of mtry = 2 seems to give the best balance between prediction accuracy and model complexity.

Country	Predicted Life Ladder 2025
Australia	7.361339
Malta	7.012267
United Kingdom	6.981340
Qatar	6.586144
Japan	6.469082
Italy	6.229648

Country	Predicted Life Ladder 2025
Estonia	6.191642
Uruguay	6.110192
Costa Rica	6.108889
Kazakhstan	5.848651

After Cross-Validating our Random Forest model, we train it on 75% of the dataset and use the remaining 25% as our test dataset to predict the Life Ladder scores for 2025. The model identifies Australia as the country with the highest predicted Life Ladder score at 7.36, followed by Malta (7.01) and the United Kingdom (6.98). These results suggest that these nations may continue to provide high overall well-being for their citizens.

Other high-ranking countries include Qatar, Japan, Italy, Estonia, Uruguay, Costa Rica, and Kazakhstan, all with relatively strong predicted Life Ladder scores. Notably, many of these countries have high GDP per capita, strong social support systems, and long healthy life expectancies, which align with our model's key predictive factors.

However, it is important to recognize that these predictions are based on historical data and assume similar trends will continue into 2025. External factors such as policy changes, economic shifts, global crises, or technological advancements could influence actual outcomes. Further refinement of the model, incorporating more dynamic features, could enhance future predictions.

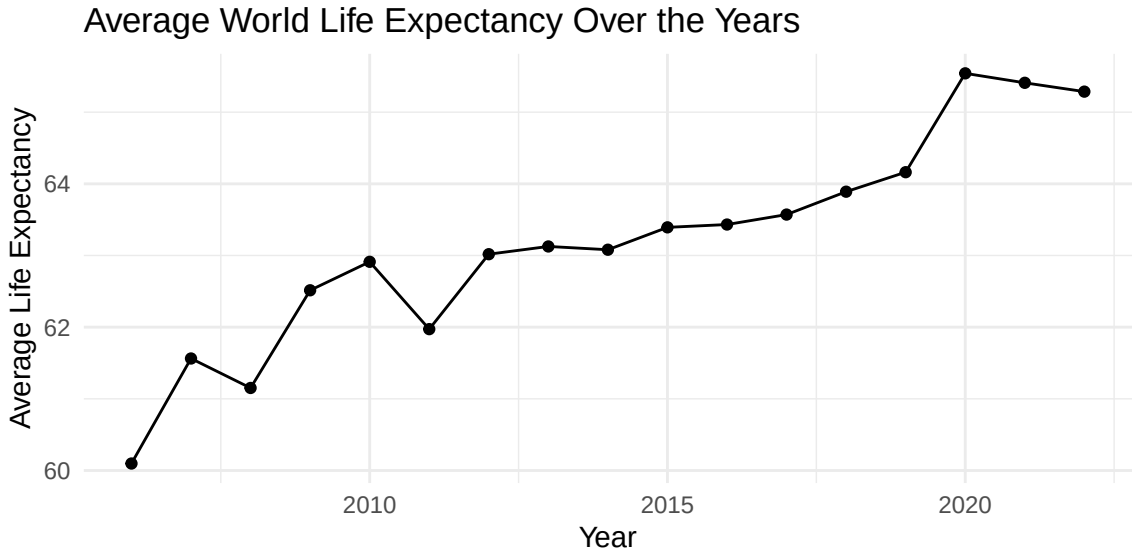
4 Life Expectancy

Life expectancy serves as a vital measure of a nation's overall health and well-being. It represents the average number of years an individual is expected to live, based on current mortality rates. This metric is significant because it offers a holistic view of living conditions, healthcare quality, lifestyle factors, and socio-economic environments within a country. Increases in life expectancy typically indicate advancements in various areas, such as better living standards, improved nutrition, enhanced education, and greater access to healthcare services, all of which contribute to a healthier and longer-lived population.

Countries with higher Gross Domestic Product (GDP) often exhibit longer life expectancies. We know this because in our previous correlation model, healthy life expectancy had the high t-value and small p-value in relation to GDP. Wealthier nations are better equipped to invest in public health initiatives and allocate resources to effectively combat diseases, resulting in healthier populations and extended lifespans. In the following analysis, we will explore global trends in life expectancy by examining its average values over time. We begin by processing the data to calculate the average life expectancy per year.

Table 6: Average Life Expectancy Over Time

Year	Average Life Expectancy
2006	60.09727
2007	61.56274
2008	61.15144
2009	62.51452
2010	62.91198
2011	61.97234
2012	63.01824
2013	63.12617
2014	63.08057
2015	63.39281
2016	63.43212
2017	63.57098
2018	63.89040
2019	64.16259
2020	65.54181
2021	65.40966
2022	65.28671



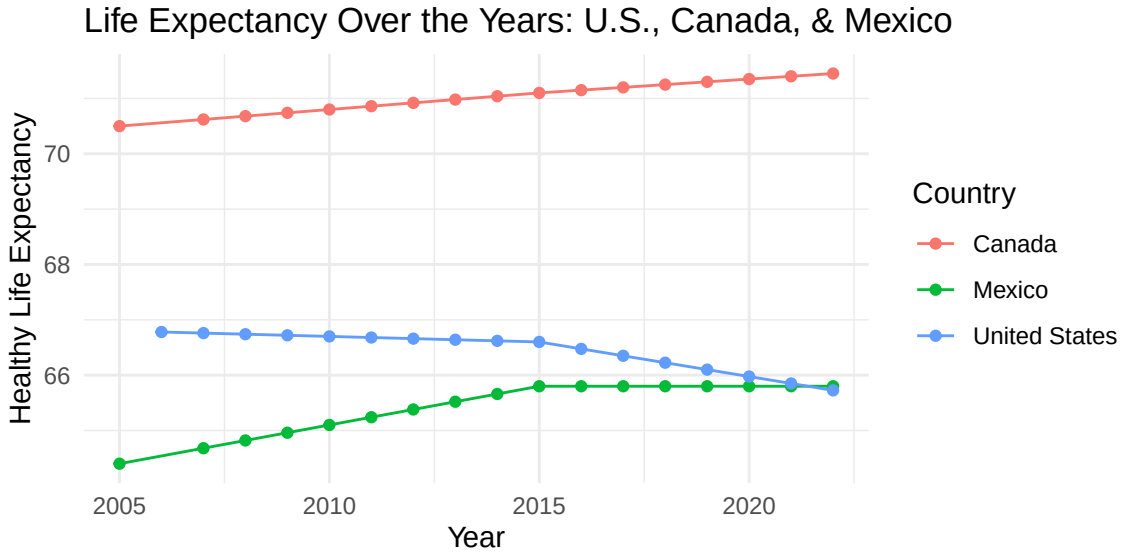
The average life expectancy worldwide from 2006 to 2022 shows a generally upward trend, though with some fluctuations. Starting at 60.10 years in 2006, life expectancy increased steadily over the years, reaching 64.16 years by 2019. This consistent growth reflects global improvements in healthcare, living conditions, and socio-economic development. However,

the data reveals minor declines in certain years, such as in 2008 and 2011, which may be attributed to regional health crises, economic challenges, or other factors that temporarily impacted mortality rates.

The year 2020 marked a significant shift, with life expectancy rising to 65.54 years, likely due to improved healthcare interventions and global health initiatives. However, the following years (2021 and 2022) saw slight declines to 65.41 years and 65.29 years, respectively. These declines may reflect the impact of the COVID-19 pandemic, which strained healthcare systems and increased mortality rates worldwide. Despite these setbacks, the overall trend from 2006 to 2022 indicates progress in global health, underscoring the importance of continued investment in healthcare infrastructure and disease prevention to sustain and improve life expectancy in the future.

4.1 U.S. v. Canada v. Mexico Life Expectancy

In this analysis, we will examine the life expectancy of three countries with strong economic and trade ties: the United States, Canada, and Mexico. The United States, a high-income country, benefits from advanced healthcare infrastructure and a diverse diet, but faces challenges such as inequality and lifestyle-related health issues. Canada, also a high-income country, has a robust universal healthcare system that contributes to a high life expectancy, with a focus on public health and preventive care. Mexico, a middle-income nation, is experiencing economic growth but still faces disparities in healthcare access and public health challenges, particularly in rural areas. By comparing these nations, we aim to uncover how factors such as healthcare systems, economic stability, and lifestyle choices influence life expectancy and health outcomes. Key insights will explore the role of healthcare policy in Canada, lifestyle and healthcare access in the U.S., and the impact of socio-economic inequalities in Mexico.



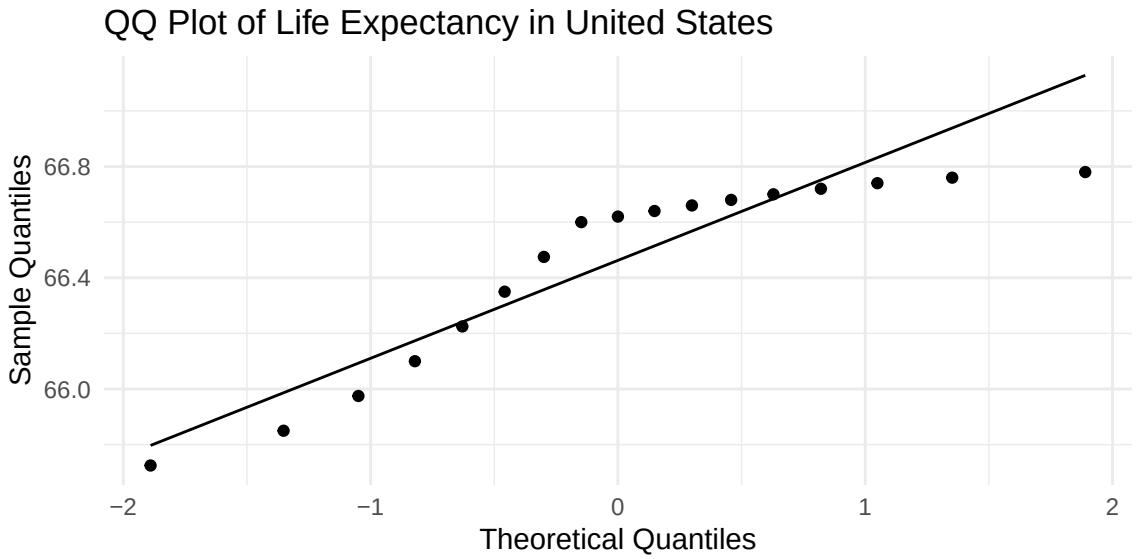
Canada maintains a consistently high and stable life expectancy throughout the period, reflecting its robust healthcare system, high standard of living, and effective public health policies. Even in 2020, Canada's life expectancy remains high, demonstrating resilience despite the COVID-19 pandemic.

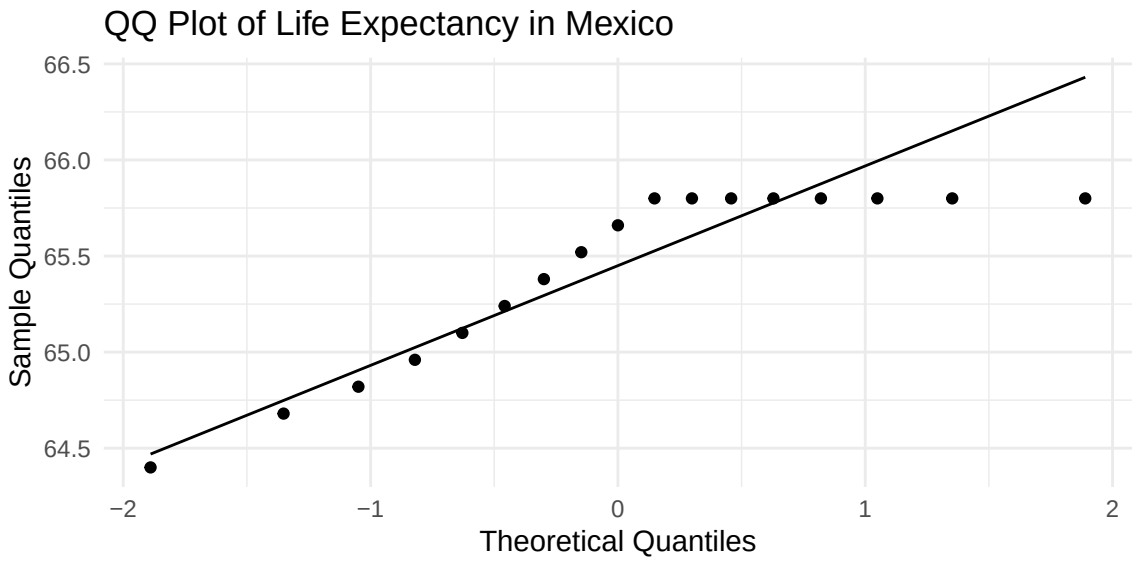
The United States shows a concerning decline in life expectancy, particularly after 2015. This decline could be largely attributed to the opioid crisis, rising obesity rates, and disparities in healthcare access. The COVID-19 pandemic further exacerbated this downward trend, leading to a significant drop in 2020.

Mexico experiences steady improvement in life expectancy from 2005 to 2020. While it starts with a lower life expectancy compared to Canada and the U.S., Mexico's consistent growth indicates progress in healthcare access and socio-economic development, with public health initiatives contributing to this positive trend. In 2020, Mexico's life expectancy continues to rise, suggesting better management of the pandemic's initial impact compared to the U.S.

Overall, Canada outperforms both the U.S. and Mexico, with its stable life expectancy reflecting the benefits of a strong healthcare system. The U.S. shows a troubling decline, while Mexico, starting lower, demonstrates steady progress. This comparison highlights the significant role healthcare systems and public health policies play in determining life expectancy, as well as the challenges faced by the U.S. in terms of public health crises and socio-economic disparities. Mexico's improvements underscore the potential for middle-income countries to achieve better health outcomes through targeted interventions and economic growth.

As we observe Mexico's steady improvement in life expectancy and the gradual decline in the U.S., the next question we aim to address is whether there is a statistically significant difference in the average life expectancy between the United States and Mexico.





The QQ plots for both the United States and Mexico show that life expectancy data are not quite normally distributed, with slight deviations at the tails. In the U.S., the deviations are more pronounced, suggesting greater variability and the presence of outliers, possibly due to events like the opioid crisis or COVID-19. Mexico's data, however, is more consistent, with fewer extreme values, reflecting its steady improvement in life expectancy.

Given that our sample size exceeds 30, as indicated by the World Happiness Report 2023, we can rely on the central limit theorem, which suggests that the sampling distribution of the mean approaches normality with larger sample sizes. This supports the use of t-tests despite the minor deviations from normality. While t-tests and ANOVA tests can yield meaningful results with large sample sizes, even if the data isn't perfectly normal, caution is necessary in interpreting these results due to the potential influence of outliers and variability.

```
    Welch Two Sample t-test

data:  us_data$healthy_life_expectancy and mexico_data$healthy_life_expectancy
t = 7.2657, df = 29.393, p-value = 4.901e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.737273 1.314492
sample estimates:
mean of x mean of y
 66.44706  65.42118
```

The results of the Welch Two-Sample t-test comparing life expectancy between the United

States and Mexico reveal a statistically significant difference. The t-value is 7.2657, indicating that the observed difference in life expectancy is substantially large compared to the variability within each group. The degrees of freedom are 29.393, reflecting the sample sizes and variability. The p-value is extremely small (4.901e-08), which provides strong evidence against the null hypothesis, suggesting a significant difference in life expectancy between the two countries. The 95% confidence interval for the difference in means ranges from 0.737273 to 1.314492, meaning that with 95% confidence, the true difference in life expectancy lies within this range. Since the confidence interval does not contain 0, this further supports the conclusion of a significant difference. The sample means show that the United States has a slightly higher life expectancy (66.45) compared to Mexico (65.42).

4.2 Comparing Positive and Negative Affect Levels between North America, Europe, and Asia

Understanding the factors that influence well-being is crucial for addressing global health disparities and improving quality of life. In this section, we analyze positive and negative affect levels across three regions—Asia, Europe, and North America—from 2005 to 2020. By examining trends in these metrics, we aim to identify regional differences and explore how cultural, economic, and social factors may contribute to variations in well-being.

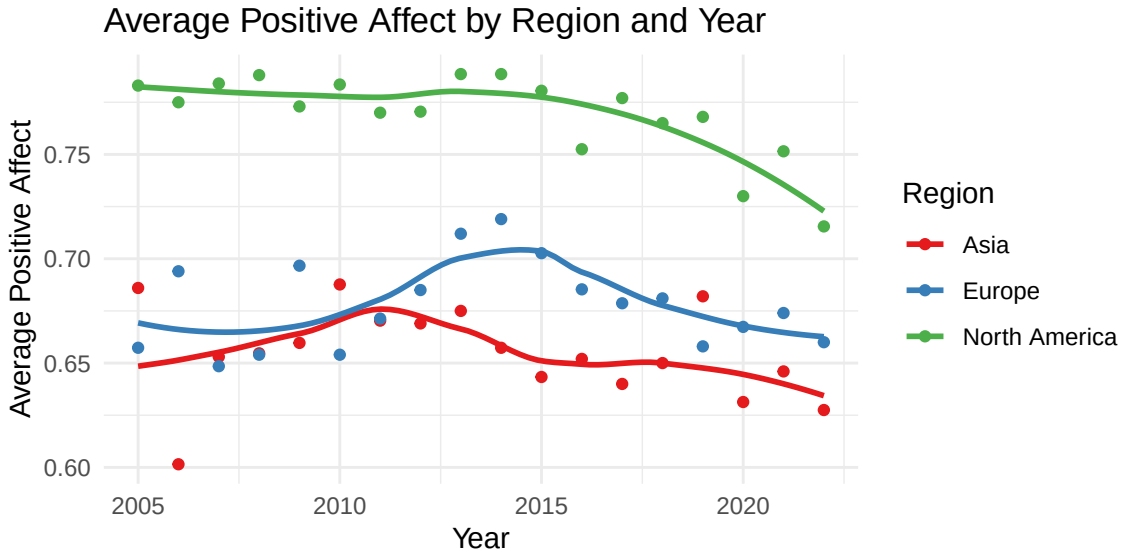


Table 7: ANOVA Table for Positive Affect by Region

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	2	0.2774894	0.1387447	57.44169	0
Residuals	132	0.3188328	0.0024154	NA	NA

In Asia, positive affect levels remain relatively stable, with only minor fluctuations, and consistently rank the highest compared to other regions. This suggests that cultural or socioeconomic factors may contribute to a sustained sense of well-being. Europe, on the other hand, shows a slight upward trend, indicating gradual improvements in positive affect over the years. North America exhibits the most pronounced increase in positive affect, with a noticeable upward trajectory from 2005 to 2020.

When comparing the regions, Asia consistently maintains the highest positive affect levels, which may be attributed to societal structures, economic stability, or cultural influences that promote well-being. North America demonstrates the most significant improvement, potentially due to increasing economic stability, enhanced social programs, or cultural shifts that prioritize well-being. Europe, while showing progress, lags slightly behind Asia and North America in terms of positive affect levels.

The ANOVA test for positive affect confirms that regional differences are statistically significant with an F-value of 57.442. The high F-value and extremely low p-value suggest that region plays a critical role in predicting positive affect levels. This result implies that cultural, economic, or social factors unique to each region significantly shape individuals' experiences of positive emotions.

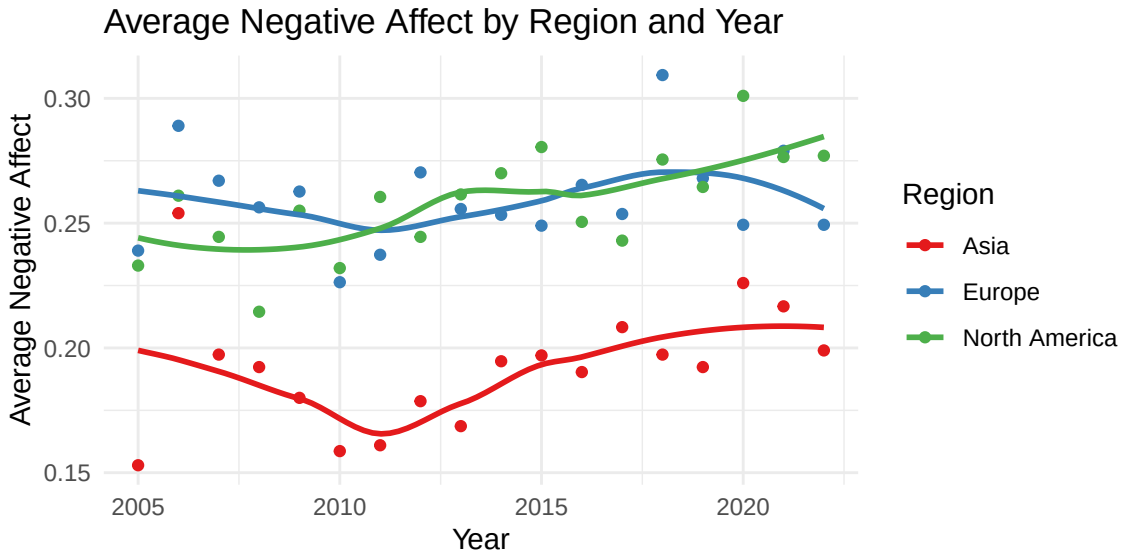


Table 8: ANOVA Table for Negative Affect by Region

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	2	0.1370405	0.0685202	33.89182	0
Residuals	132	0.2668688	0.0020217	NA	NA

The trends in negative affect levels across regions also highlight important patterns. Asia maintains a relatively stable trend in negative affect, with minor fluctuations, and consistently records the lowest negative affect levels compared to other regions. This suggests that cultural or socio-economic structures in Asia may contribute to lower stress and fewer negative emotions. Europe displays a slight downward trend in negative affect, indicating gradual reductions in stress and negative emotions over time. North America, similar to its trend in positive affect, exhibits the most pronounced improvement, with a significant decline in negative affect from 2005 to 2020.

When comparing the regions, Asia consistently reports the lowest levels of negative affect, reinforcing the idea that cultural or societal factors help mitigate stress and negative emotions. North America shows the most substantial reduction in negative affect, possibly due to increasing access to mental health resources, economic stability, or shifts in cultural attitudes toward well-being. Europe, while making progress, lags slightly behind Asia and North America in reducing negative affect levels.

The ANOVA test for negative affect reveals a statistically significant difference across regions with an F-value of 33.892. The high F-value and extremely low p-value confirm that regional differences play a crucial role in predicting negative affect levels. This finding suggests that cultural, economic, and social influences have a strong impact on the prevalence of negative emotions across different regions.

5 Conclusion

Our analysis reveals several important insights into the factors influencing happiness across different regions. Generally, GDP per capita is positively correlated with well-being indicators such as life ladder, social support, and healthy life expectancy, while higher GDP is associated with lower perceptions of corruption and reduced generosity. Our Random Forest model for predicting the top 10 happiest countries in 2025 performed well, with an optimal mtry of 2, yielding an RMSE of 0.3732 and an R-squared value of 0.8570. The model predicted Australia, Malta, and the United Kingdom as the top three countries, based on factors like high GDP, social support, and life expectancy. Regional trends in positive and negative affect show that Asia consistently maintains higher positive affect and lower negative affect compared to Europe and North America, with the latter regions showing gradual improvements. A comparison of life expectancy trends between China and the United States highlights a significant gap, as China has seen steady increases in life expectancy, while the U.S. has faced health challenges that have limited improvements. Furthermore, the significant differences in negative affect levels across regions, as indicated by ANOVA tests, suggest that cultural, economic, and political factors heavily influence emotional well-being. These findings underscore the complex relationship between economic conditions, social support, and health, emphasizing the need for targeted policies to address disparities and enhance global well-being.