Final Project

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

Jackson Bolcer (jacksonbolcer) Navin Lo (navin)
Yuval Hod (yuvalhod) Joey Peir (jpeir)

Abstract

In this report we are investigating the relationship between economic factors, particularly GDP per capita, and happiness levels across various countries using data from the World Happiness Report (2008-present). Our objective is to understand how economic well being influences the happiness index, represented by the Life Ladder score. Through exploratory data analysis, correlation analysis, and multivariate regression modeling, the report aims to provide insights into the role of economic policies in terms happiness. Additionally, the analysis considers other contributing factors such as social support, health, freedom to make life choices, generosity, and perceptions of corruption, offering a view of the different aspects that contribute to happiness.

Introduction

The World Happiness Report is an annual publication that measures and ranks countries based on their levels of happiness and well being. The dataset utilized in this report spans from 2008 to the present and includes a variety of indices related to happiness. The focus of this report is to examine the influence of economic factors, particularly GDP per capita, on the happiness levels across different countries.

The dataset we are using is comprised of several key variables which will help us answer our question:

1. `Country name`: The name of the country.
2. `Year`: The year of the observation.
3. `Life Ladder`: The primary happiness index, representing overall life satisfaction.
4. `Log GDP per capita`: A measure of economic performance which is adjusted for the population size.
5. `Social Support`: The availability of social networks and support systems.
6. `Healthy life expectancy at birth`: A measure of the population's health.
7. `Freedom to make life choices`: The degree of individual freedom in making life decisions.
8. `Generosity`: The willingness of individuals to help others.
9. `Perceptions of corruption`: The perceived level of corruption in the country.
10. `Positive affect`: The frequency of positive emotions experienced.
11. `Negative affect`: The frequency of negative emotions experienced.

This report aims to answer the following key question: How do economic factors influence happiness levels across different countries? By investigating this question, our report seeks to provide insights into the importance of economic wellbeing in enhancing happiness while identifying significant factors contributing to national happiness levels across the spectrum. Our analysis includes exploratory data visualization to understand distributions and relationships, correlation analysis to examine direct relationships, and regression modeling to quantify the impact of economic and other factors on happiness. Our report aims to inform policy decisions aimed at improving overall well being and happiness.

Data

Step 1: Loading and Exploring the Data

Lets take a look at the data:

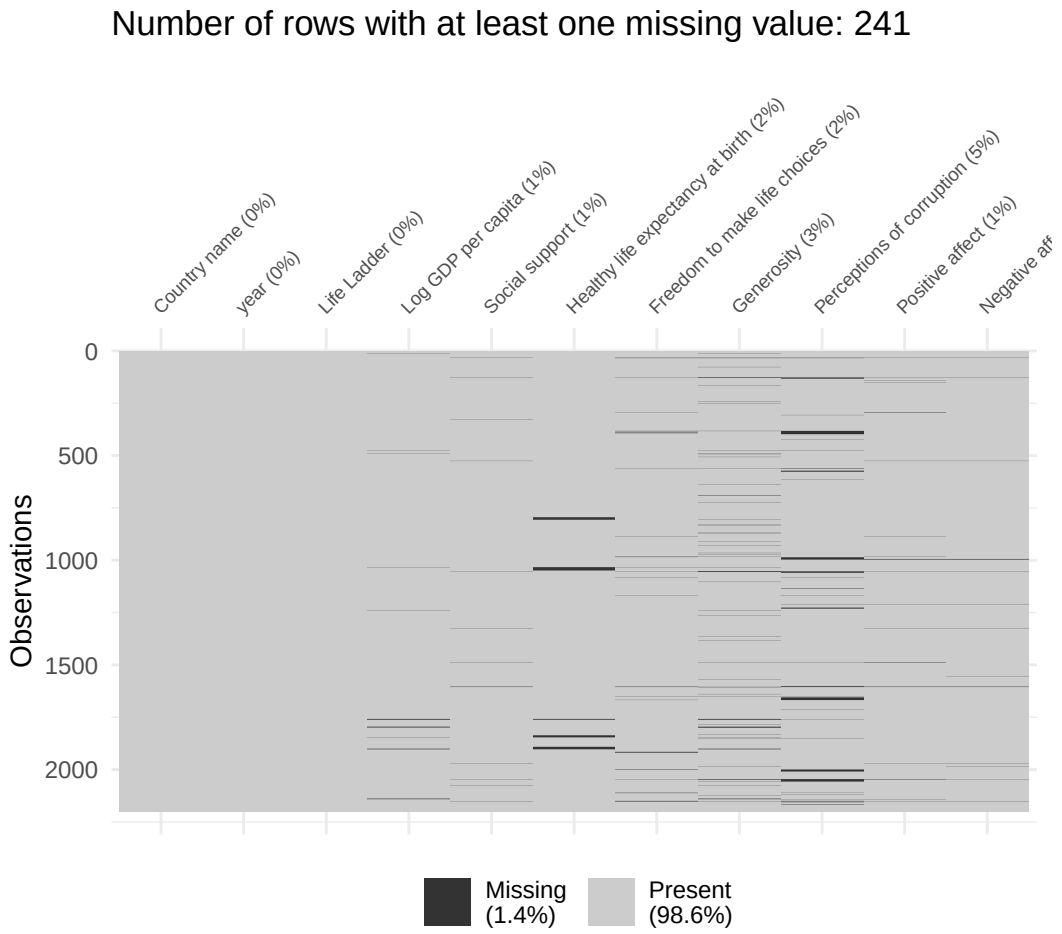
```
  Country name year Life Ladder Log GDP per capita Social support
1  Afghanistan 2008       3.724              7.350          0.451
2  Afghanistan 2009       4.402              7.509          0.552
3  Afghanistan 2010       4.758              7.614          0.539
4  Afghanistan 2011       3.832              7.581          0.521
5  Afghanistan 2012       3.783              7.661          0.521
  Healthy life expectancy at birth Freedom to make life choices Generosity
1                             50.5                         0.718      0.168
2                             50.8                         0.679      0.191
3                             51.1                         0.600      0.121
4                             51.4                         0.496      0.164
5                             51.7                         0.531      0.238
  Perceptions of corruption Positive affect Negative affect
1                     0.882           0.414           0.258
2                     0.850           0.481           0.237
3                     0.707           0.517           0.275
4                     0.731           0.480           0.267
5                     0.776           0.614           0.268
```

```
[1] 2199   11
```

We can see that we have 2199 observations in our data that covers 11 variables as expected.

Missing Data

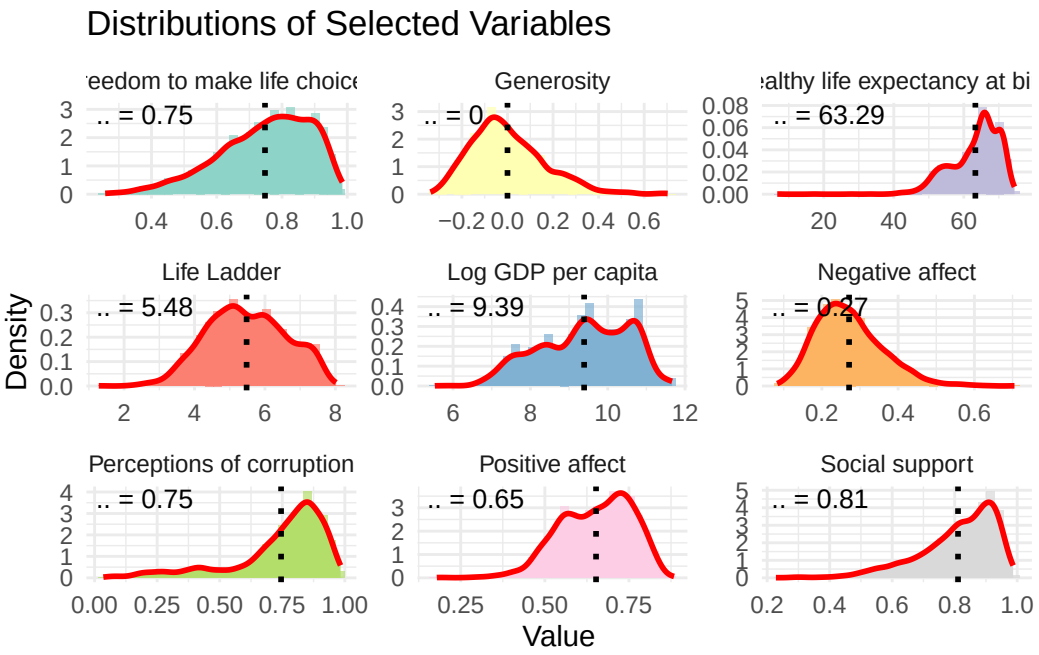
Next, let's see where the missing values of this dataset come from:



It can be seen in this plot Generosity and Perceptions of Corruption make up most of the missing values, and that missing values span over 241 rows, so it would not be smart to remove them.

Exploring Distributions

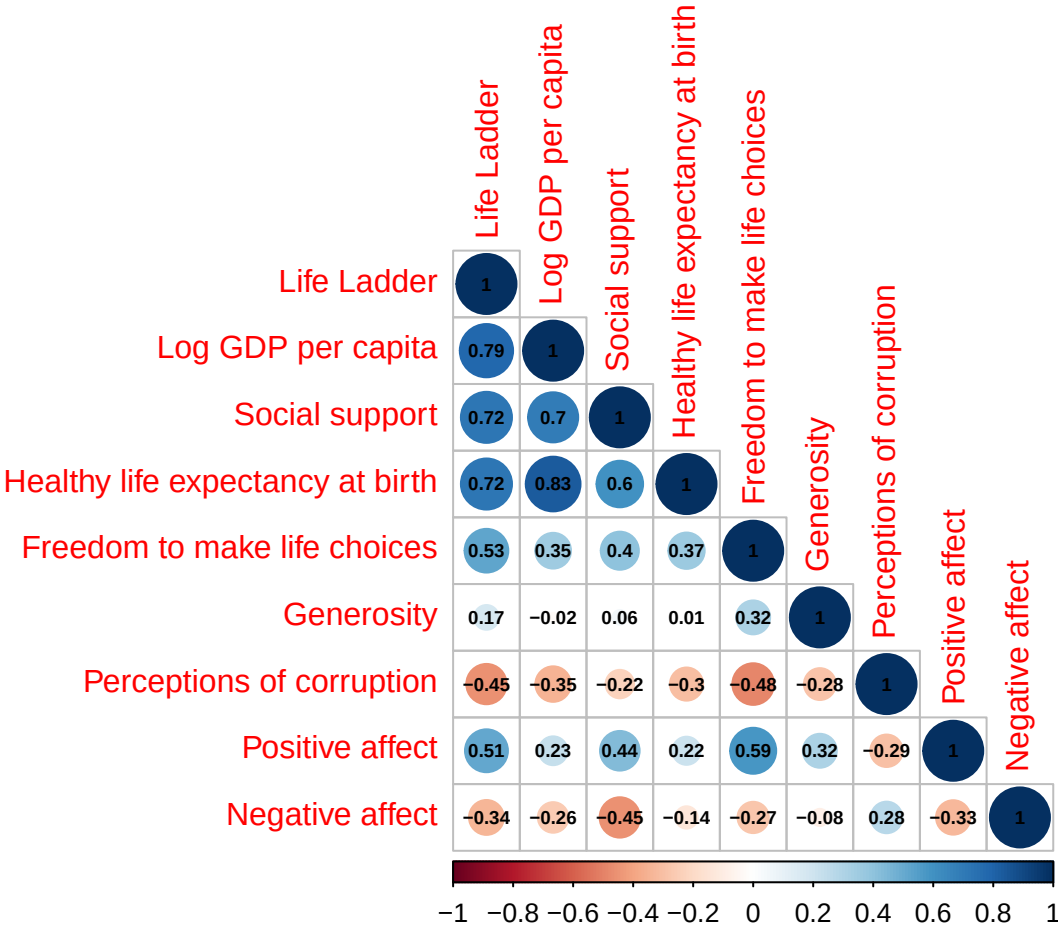
Now, let's take a look at the distribution of some of our variables.



From the above plots, it is very clear that only `Life Ladder` closely resembles a normal distribution, while all other columns are skewed heavily one way or the other. We will have to transform those variables.

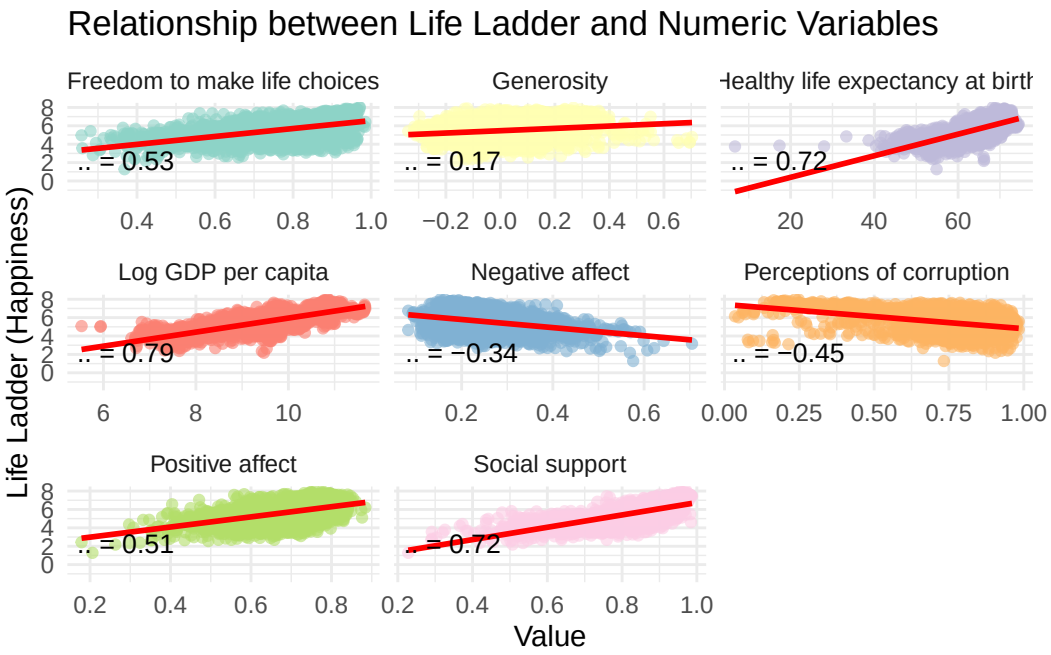
Exploring Numerical Variable Correlation

Next, lets look at the correlation between each of the variables in our dataset. With this matrix we will be able to see what variables are most closely correlated to our outcome variable Happiness (Life Ladder). We can also use this plot to analyze the possibility of colinearity between predictor variables.



When we look at the second column for `Life Ladder`, it can be seen that `Log GPD per capita` is the most correlated with a value of 0.79. We changed `year` to a categorical variable.

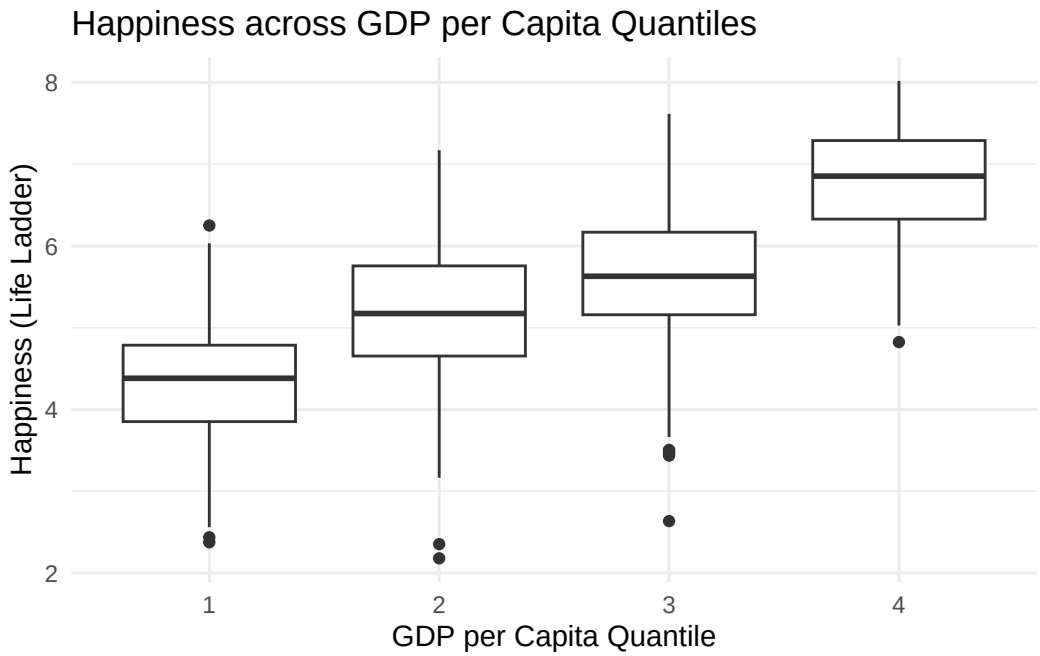
Below is a visualization of the variable `Life Ladder` in correlation to the numeric variables in the dataset:



Above, we decided to look at all of the numeric variables in relation to happiness in hopes of getting steps closer to reaching our answer. As we examine the results of the graphs we have, especially in graph 4, we are able to see a direct relationship between happiness (Life Ladder) and having more money (`Log GPD per Capita`), as the line of best fit has a positive slope and the ρ value is the highest. It is important to note that even though this seems to be a direct relation, there are more outliers as we examine observations with a higher `Log GDP per Capita`. These graphs validate our claims from the correlation matrix. It is also important to note that `Negative Affect` and `Perceptions of Corruption` are negatively correlated to Happiness.

Exploring Categorical Variables

The next visualizations below are for understanding the relationship between our categorical variables (`Year`, `GDP_quantile`, and `Country Name`) and our outcome variable (`Life Ladder`). Let's examine a boxplot of `Log GDP per capita` in four quantiles to see how happiness is distributed across quantiles.



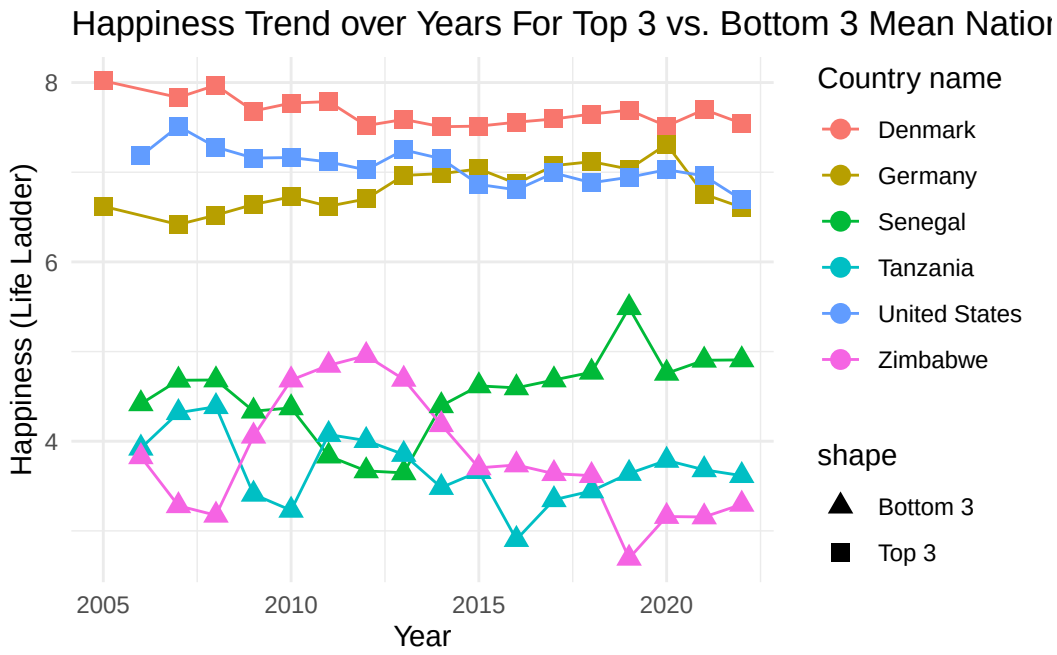
Clearly it can be observed that as quantile of Log GDP increases, Happiness also increases.

As a way of picking countries, we are interested in the top 3 countries with the highest `Log GPD per capita` mean values across years as well as the bottom 3 countries with this lowest value to examine happiness levels. To examine this relationship we are going to analyze how their happiness has changed over time with respect to this variable. Let's see what the top and bottom 3 countries are for our `Log GDP per capita` variable:

```
# A tibble: 3 x 2
  `Country name` `Mean GDP (Top 3)`
  <chr>                       <dbl>
1 United States                11.0
2 Denmark                      10.9
3 Germany                      10.8
```

```
# A tibble: 3 x 2
  `Country name` `Mean GDP (Bottom 3)`
  <chr>                           <dbl>
1 Zimbabwe                         7.61
2 Tanzania                         7.70
3 Senegal                          8.02
```

Now that we have the top and bottom 3 countries with average `Log GPD per capita`, lets examine how happiness changes over time in an effort to analyze a trend in the data.



As we can see in the above time series plot, there is a very clear level difference in happiness levels between our selected low and high GDP countries (Triangle vs. Square). Additionally it can be seen over time that the level of happiness in the higher averaged GDP countries is much more constant as opposed to the more volatile level of happiness in the lower averaged GDP countries.

Step 3: Fitting Regression Models

Simple Regression Model

Below we are going to be fitting a regression model with `Log GDP per capita` as the independent variable and `Life Ladder` as the dependent variable:

```
Call:
lm(formula = `Life Ladder` ~ `Log GDP per capita`, data = whr_2023)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3679 -0.4745 -0.0117  0.5071  2.5449

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.68302    0.12220  -13.77   <2e-16 ***
`Log GDP per capita`  0.76337    0.01292   59.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6953 on 2177 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.616, Adjusted R-squared:  0.6158
```

```
F-statistic:  3493 on 1 and 2177 DF,  p-value: < 2.2e-16
```

Breaking down the summary from above, we will first examine the residuals which provide a summary of the difference between the observed values and the values predicted by the model. Since the median residual are close to zero, this suggests that the model predictions are fairly accurate. Now looking at the intercept, -1.68302 is the expected value of the Life Ladder score when `Log GDP per capita` is zero. While this value might not have a direct real-world interpretation since `Log GDP per capita` would not realistically be zero, it serves as a baseline for the model we created. Now looking at the `Log GDP per capita` 0.76337 which indicates that for every one unit increase in `Log GDP per capita`, the `Life Ladder` score increases by approximately 0.76337 units. This positive coefficient we examine suggests a positive relationship between `Log GDP per capita` and happiness. Moving along to the p-value we can assess that the low p-value indicates that the relationship between `Log GDP per capita` and the `Life Ladder` score are highly significant. Interpreting the R^2 value of 0.616, this indicates that approximately 61.6% of the variance in the `Life Ladder` score can be explained by `Log GDP per capita` while the adjusted R^2 value of 0.6158 which is very close to the R^2 value indicated that the model is a good fit. To conclude, the regression analysis indicates a strong, positive, and statistically significant relationship between `Log GDP per capita` and the `Life Ladder` score. The model explains a substantial portion of the variance in happiness levels across countries, suggesting that economic factors play a crucial role in determining national happiness.

Step 4: Analyzing the Relation between Log GDP per Capita and Happiness (Life Ladder)

Multiple Regression Model

After looking at the results above, it is important to understand the relationship between `Log GDP per capita` and happiness(`Life ladder`) while accounting for other potential influential factors. This will help in isolating the effect of `Log GDP per capita` on happiness and ensures that the observed relationship is not shaken up by other variables.

```
Call:
lm(formula = `Life Ladder` ~ `Log GDP per capita` + `Social support` +
    `Healthy life expectancy at birth` + `Freedom to make life choices` +
    Generosity + `Perceptions of corruption` + `Positive affect` +
    `Negative affect`, data = whr_2023)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7976 -0.3102  0.0355  0.3267  1.8172

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -2.73620    0.17573 -15.571  < 2e-16 ***
`Log GDP per capita`               0.39753    0.02175  18.278  < 2e-16 ***
`Social support`                   1.81996    0.16109  11.298  < 2e-16 ***
`Healthy life expectancy at birth` 0.02756    0.00318   8.667  < 2e-16 ***
`Freedom to make life choices`     0.37857    0.12062   3.139  0.00172 **
```

```
Generosity                          0.35070    0.08312   4.219 2.56e-05 ***
`Perceptions of corruption`        -0.69708    0.08066  -8.643  < 2e-16 ***
`Positive affect`                   2.31304    0.15099  15.319  < 2e-16 ***
`Negative affect`                  -0.01465    0.16879  -0.087  0.93082
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5365 on 1949 degrees of freedom
  (241 observations deleted due to missingness)
Multiple R-squared:  0.7799,    Adjusted R-squared:  0.779
F-statistic:   863 on 8 and 1949 DF,  p-value: < 2.2e-16
```

	Coefficient	P.Value
(Intercept)	-2.73619858	1.297371e-51
`Log GDP per capita`	0.39752564	5.163895e-69
`Social support`	1.81996131	1.031523e-28
`Healthy life expectancy at birth`	0.02756492	9.136532e-18
`Freedom to make life choices`	0.37857463	1.723286e-03
Generosity	0.35069837	2.564623e-05
`Perceptions of corruption`	-0.69707998	1.123410e-17
`Positive affect`	2.31304147	4.220518e-50
`Negative affect`	-0.01465482	9.308209e-01

As we analyse the relationship between these two variables of happiness and GDP per capita, it is important for us to consider other variables that might influence the dependent variable since a variable like the one we are assessing(happiness - Life Ladder score) could be affected by multiple factors. If we brush past these variables our analysis might be wrong and might fall into biases which we do not see ourselves.

Analyzing the Results of Our Data:

Breaking down the summary from the multivariate regression model, we first examine the residuals. In this case the median residual is close to zero, suggesting that the model predictions are fairly accurate. Now looking at the intercept, -2.73620 which serves as a baseline for the model. Examining the coefficient for `Log GDP per capita`, 0.39753 indicates that for every one unit increase in `Log GDP per capita`, the `Life Ladder` score increases by approximately 0.39753 units, holding all other variables constant. This positive coefficient suggests a positive relationship between log GDP per capita and happiness. The very low p-value ($< 2.2e^{-16}$) for this coefficient indicates that the relationship between `Log GDP per capita` and the `Life Ladder` score is highly significant. The coefficient for `Social support`, 1.81996, indicates that higher social support is associated with higher happiness levels, holding all other factors constant. The low p-value ($< 2.2e^{-16}$) confirms that this relationship is statistically significant. For `Healthy life expectancy at birth`, the coefficient is 0.02756, suggesting that better health outcomes are associated with higher happiness, holding other factors constant. Again, the very low p-value ($< 2.2e^{-16}$) indicates a statistically significant relationship.The coefficient for `Freedom to make life choices` is 0.37857, suggesting that greater freedom to make life choices is associated with higher happiness, holding other factors constant. The p-value (0.00172) indicates that this relationship is statistically significant. `Generosity` has a coefficient of 0.35070, suggesting

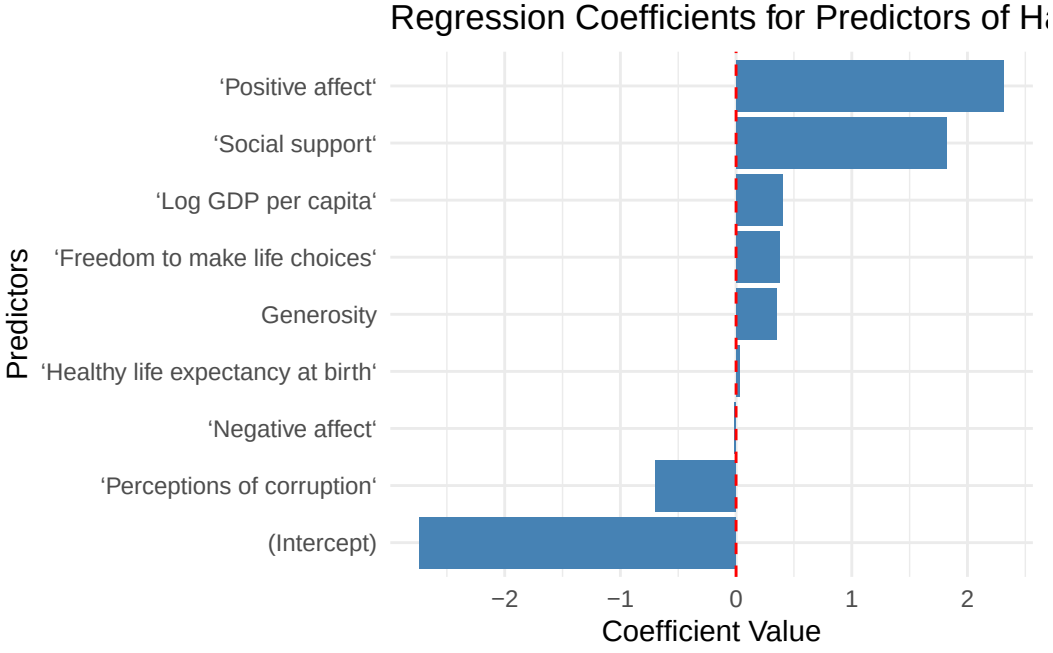
that higher levels of generosity are associated with higher happiness, holding other factors constant. The p-value ($2.56e-05$) confirms that this relationship is statistically significant.The coefficient for `Perceptions of corruption` is -0.69708, indicating that higher perceptions of corruption are associated with lower happiness levels, holding other factors constant. The very low p-value ($< 2.2e^{-16}$) indicates a statistically significant relationship. `Positive affect` has a coefficient of 2.31304, indicating that more frequent positive emotions are associated with higher happiness, holding other factors constant. The very low p-value ($< 2.2e^{-16}$) indicates a statistically significant relationship. Lastly, the coefficient for `Negative affect` is -0.01465, suggesting a very weak relationship between negative emotions and happiness when other factors are controlled. The high p-value (0.93082) indicates that this relationship is NOT statistically significant.

Observing the R^2 value of 0.7799 indicates that approximately 77.99% of the variance in the `Life Ladder` score can be explained by the predictors in the model. The adjusted R^2 value of 0.779, which is very close to the R^2, suggests that the model is a good fit. The F-statistic of 863 on 8 and 1949 degrees of freedom, with a p-value $< 2.2e^{-16}$, tests the overall significance of the model and indicates that the model is statistically significant.

To conclude, the multivariate regression analysis indicates that several factors, including `Log GDP per capita`, `social support`, `healthy life expectancy`, `freedom to make life choices`, `generosity`, `perceptions of corruption`, and `positive affect` have significant relationships with the `Life Ladder` score. The model explains a substantial portion of the variance in happiness levels across countries, suggesting that both economic and non-economic factors play crucial roles in determining national happiness.

Step 5: Visualizing Regression results

Now we will visualize the regression results to understand how each variable relates to happiness:



The variables that have the highest coefficient values are `Positive Affect` and `Social Support`, and the variable with the lowest coefficient value is `Perceptions of Corruption`. If we think about the

distributions of each variable, we saw earlier that the means of `Positive Affect` and `social support` were less than 1. Therefore, a one unit increase in these variables would then cause a drastic increase in `Life Ladder`. This plot could be misleading about the importance of each variable.

Scaling Data to Better Understand Variables

As mentioned above, the variables are not scaled so it can be misleading. We went ahead and scaled the data so that we could get a better understanding of the importance of each variable in relation to `Life Ladder`.

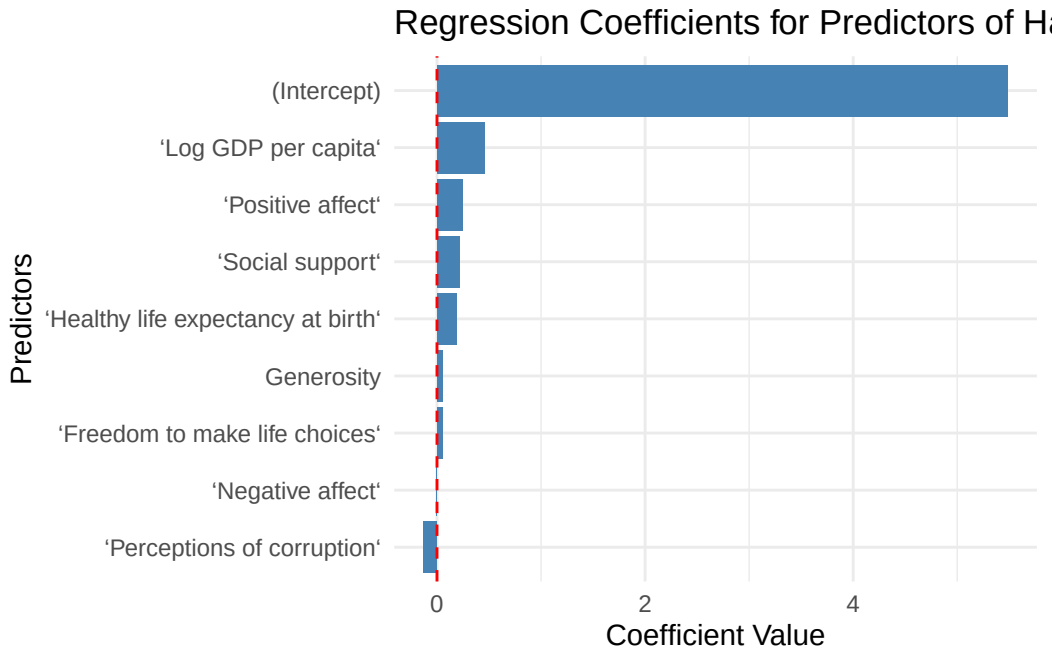
```
Call:
lm(formula = `Life Ladder` ~ `Log GDP per capita` + `Social support` +
    `Healthy life expectancy at birth` + `Freedom to make life choices` +
    Generosity + `Perceptions of corruption` + `Positive affect` +
    `Negative affect`, data = scaled_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7976 -0.3102  0.0355  0.3267  1.8172

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5.484734   0.012144 451.643  < 2e-16 ***
`Log GDP per capita`              0.458507   0.025085  18.278  < 2e-16 ***
`Social support`                  0.220129   0.019485  11.298  < 2e-16 ***
`Healthy life expectancy at birth` 0.190228   0.021948   8.667  < 2e-16 ***
`Freedom to make life choices`    0.053052   0.016904   3.139  0.00172 **
Generosity                        0.056490   0.013389   4.219 2.56e-05 ***
`Perceptions of corruption`      -0.129542   0.014988  -8.643  < 2e-16 ***
`Positive affect`                 0.244980   0.015992  15.319  < 2e-16 ***
`Negative affect`                -0.001273   0.014663  -0.087  0.93082
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5365 on 1949 degrees of freedom
  (241 observations deleted due to missingness)
Multiple R-squared:  0.7799,    Adjusted R-squared:  0.779
F-statistic:   863 on 8 and 1949 DF,  p-value: < 2.2e-16
```

Notice how the only values that changed from the last linear model to this scaled-data model is the Estimate values. All of the P-values remained the same.

Lets once again plot the Regression coefficients for predictors of Happiness with our new scaled data:

```
                                  Coefficient      P.Value
(Intercept)                       5.484734134 0.000000e+00
`Log GDP per capita`              0.458506868 5.163895e-69
```

```
`Social support`                     0.220129288 1.031523e-28
`Healthy life expectancy at birth`   0.190228377 9.136532e-18
`Freedom to make life choices`       0.053052481 1.723286e-03
Generosity                           0.056490151 2.564623e-05
`Perceptions of corruption`         -0.129541752 1.123410e-17
`Positive affect`                    0.244980148 4.220518e-50
`Negative affect`                   -0.001273087 9.308209e-01
```



Ignoring the intercept, we can now see that the `Log GDP per capita` variable has the largest impact on increasing Happiness. It is still important to note that `Positive Affect` and `Social Support` still have a significant positive impact on the outcome variable, just not nearly as prominent after scaling the data.

Conclusion

In this project, we aimed to investigate the relationship between economic factors, particularly GDP per capita, and happiness levels across various countries using data from the World Happiness Report spanning from 2008 to the present. Our objective was to understand how economic well-being influences the happiness index, represented by the Life Ladder score, while also considering other contributing factors such as social support, health, freedom to make life choices, generosity, and perceptions of corruption.

We began by loading and exploring the data, examining distributions, and identifying missing values. Through exploratory data analysis, we visualized the distributions of key variables and their correlations with happiness. Our analysis showed that `Log GDP per capita` had a strong positive correlation with happiness (`Life Ladder`), and other factors like social support, healthy life expectancy, and freedom to make life choices also played significant roles.

We then fitted both simple and multiple regression models to quantify the impact of these factors on happiness. The simple regression model highlighted a significant positive relationship between Log GDP per capita and happiness, explaining about 61.6% of the variance in happiness levels. The multiple regression model, which included additional predictors, further reinforced the importance of economic factors while also highlighting the significant contributions of social support, health, freedom, generosity, and perceptions of corruption. This model explained approximately 77.99% of the variance in happiness levels.

Visualizing the regression results with both original and scaled data allowed us to understand the relative importance of each predictor. Scaling the data revealed that Log GDP per capita had the largest impact on increasing happiness, followed by positive affect and social support.

In conclusion, our analysis confirms that economic well-being, as measured by Log GDP per capita, is a crucial determinant of national happiness. However, non-economic factors such as social support, health, freedom to make life choices, generosity, and perceptions of corruption also significantly influence happiness levels. These findings suggest that while economic policies aimed at improving GDP per capita are essential, other non-ecomomic approaches that enhance social support, health, freedom, and reduce corruption are equally important in maintaining national happiness and well-being. This comprehensive understanding can inform policy decisions aimed at improving overall well-being and happiness.