ISyE 312

Group 14

Chris Duffy, Ian Leonhardt, Olivia Vitale, Ryan Susilo

Unhappy With Where You Live? Here’s What You Should Consider

1. **Introduction**

The *World Happiness Report* portrays the level of happiness in countries throughout the world and helps to explain personal and national variations in happiness through surveys. The first report was conducted in 2005 and our dataset utilizes all the data up until 2020. We also utilized the 2021 report which is separate and contains a regional indicator. This landmark survey continues to gain global popularity among governments, organizations, and civil societies who reference happiness indicators in efforts to better their policies. The happiness scores utilize data from the Gallup World Poll. The survey is conducted by asking the participants to rate their current lives on a scale from 0 to 10 by imagining a ladder, hence the name in the dataset itself. Six factors are included in the dataset regarding the happiness score: GDP, social support, life expectancy, freedom, generosity, and corruption. The goal of the factors is not to impact the total score, but to show how each of them explain the overall happiness score. The research we conducted was focalized on how this data could be utilized when helping advise someone on what factors they should consider before moving to a different country

1. **Exploring the Dataset**

The *World Happiness Report* dataset contains 8 variables and 1,949 rows. The variables are as follows:

Country: Country where the data was polled from

Year: The year the data was collected

GDP: Gross domestic product, or the economic production

Social Support: A support system in times of need/crisis

Life Expectancy: Average lifespan of people

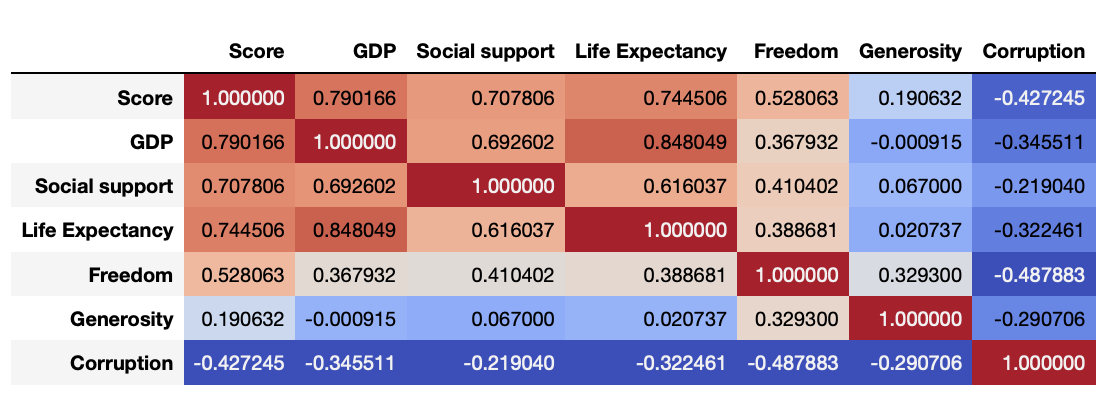
Freedom to Make Life Choices: Presence of free speech, actions, beliefs, etc...

Generosity: Response to “Have you recently donated to charity?”

Perceptions of Corruption: Within the government, state and local, and corporations

The target is trying to predict the happiness score (or “Life Ladder”), which is a continuous, moving value. The variables GDP, social support, life expectancy, freedom to make life choices, generosity, and perceptions of corruption are all continuous. The variable country is categorical and year is discrete.

We began our research by first observing the correlation between the six factors that explain the overall happiness score and the happiness score itself.



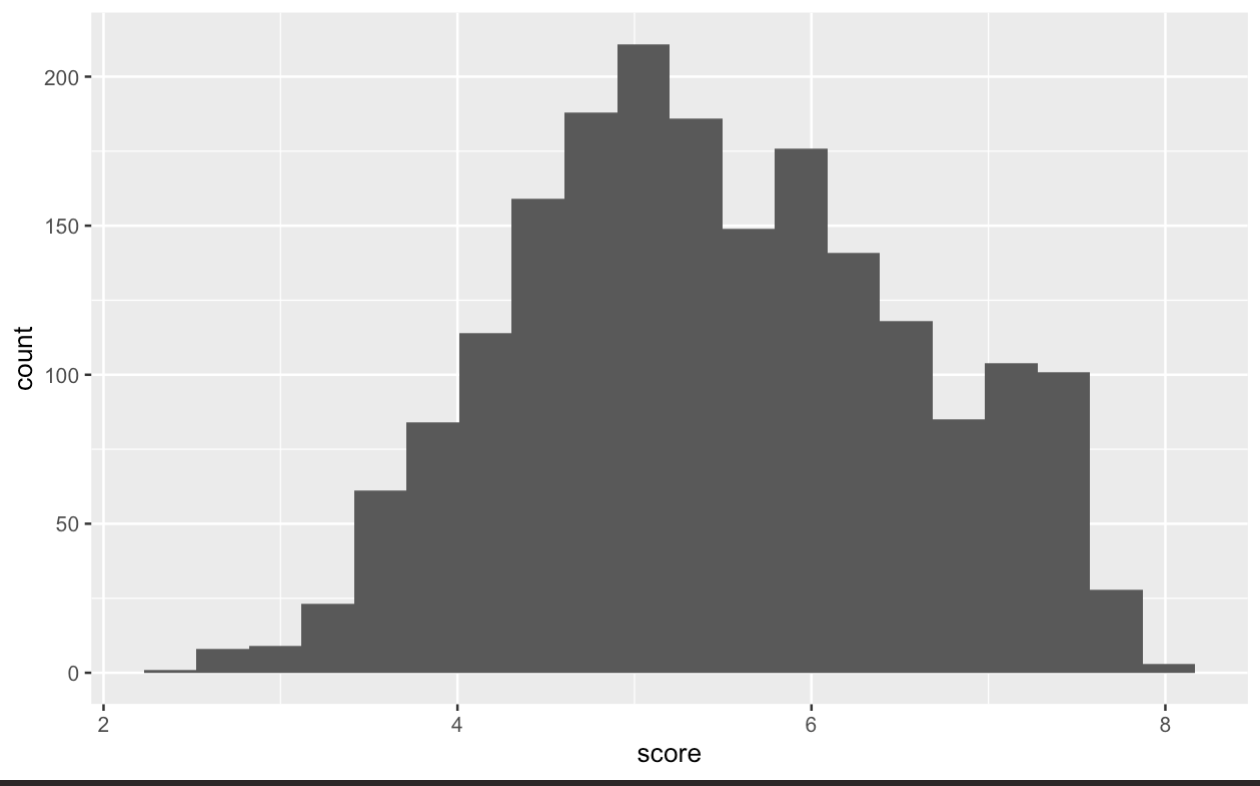
From this analysis we concluded that GDP, social support, and life expectancy weigh the most on the happiness score and decided to focus our attention there.

1. **Objective**

Happiness is often a factor that isn’t considered when talking about different countries, however it is arguably one of the most important aspects in life to have. The research we conducted was focalized on how this data could be utilized when helping advise someone on where they should move to. In addition, the goal of our research was to be able to put together a report explaining the trends in world happiness, ultimately serving as a resource to someone who may be unhappy with their current state of living and is seeking a move. In order to do this, we tried to answer the following 3 questions:

1. Does the economic stability of a country (GDP) have an effect on the happiness of those living there?
2. How does the level of social support found within a country contribute to its happiness?
3. How does a country’s life expectancy affect its happiness? What does this mean on a deeper level?
4. **Approach to answer the identified questions**

Our approach to answering these answers was conducted through the use of RStudio and Python. Multiple different types of data analysis were conducted such as making visualizations of the data correlations and running simple linear regression models as well as multiple linear regression models. In order to begin our approach to answer these questions, we thought it would be useful to do some initial exploratory analysis to get familiar with the data.

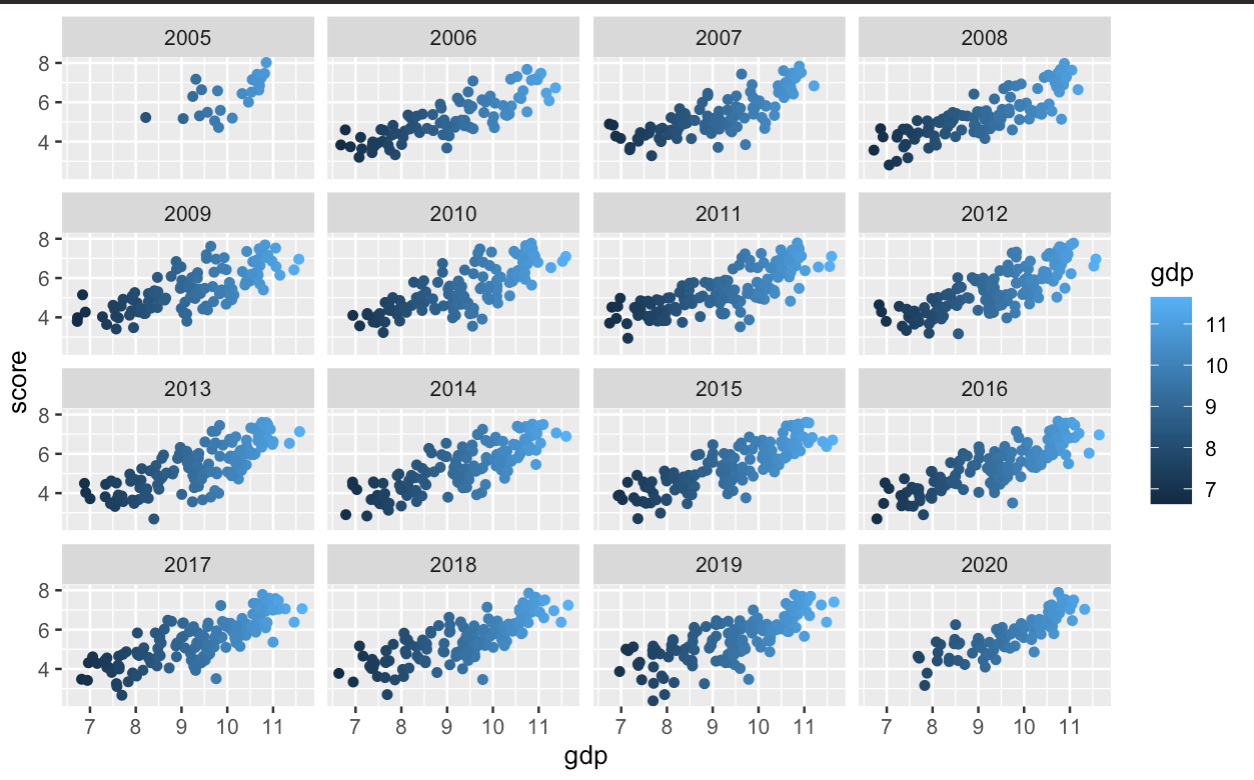


It can be seen from this histogram, which used a bin width of 20, that seldom does a country score 7 and above. In addition, a country will almost never score below a 3 on the 0-8 scale. This allows us to gain more insight into the overall happiness levels of the globe as well as why some regression models might be performing in a particular manner.

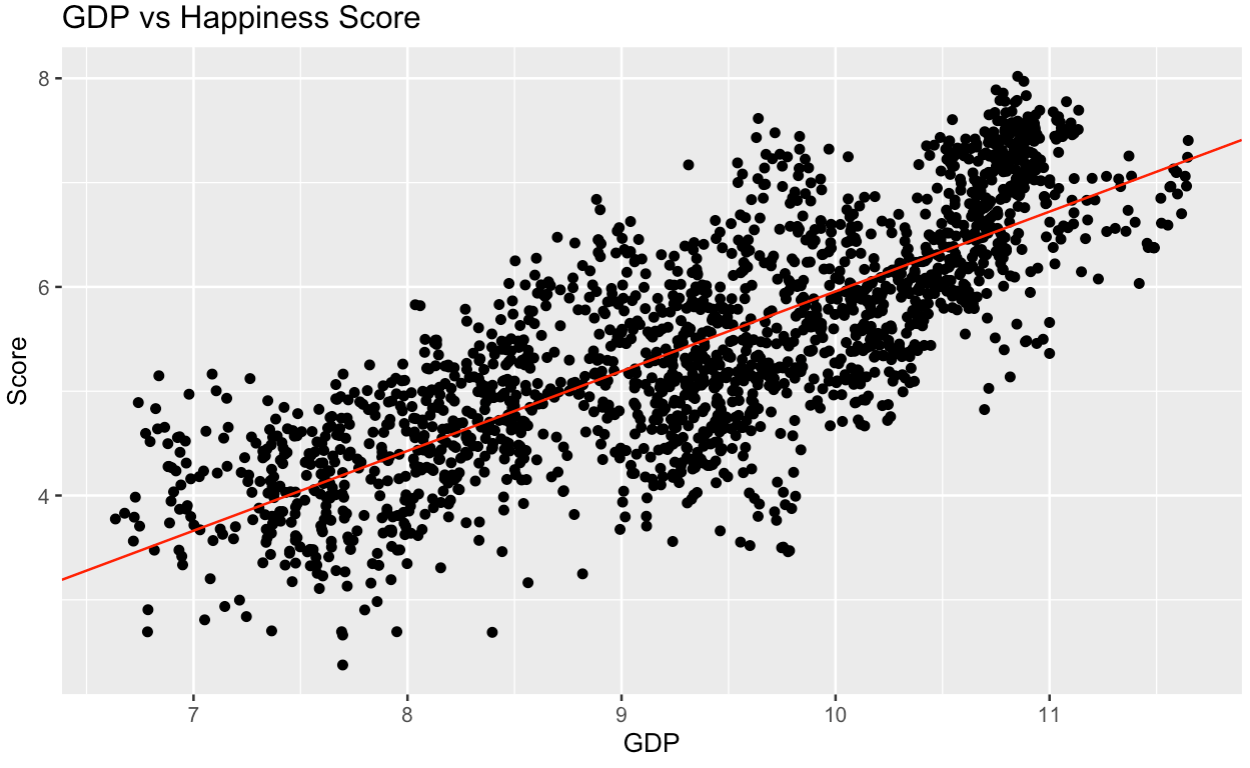
* 1. **Economic Stability’s (GDP) effect on Overall Happiness**

When considering the overall happiness, there is a certain notion or stereotype that money buys happiness and that it will fix anyone’s problems. Analyzing the effect of a country’s economic stability on its overall happiness is a way in which this theory can be looked into.

We started off by looking at GDP vs Happiness scores in a scatter plot form over the past years.

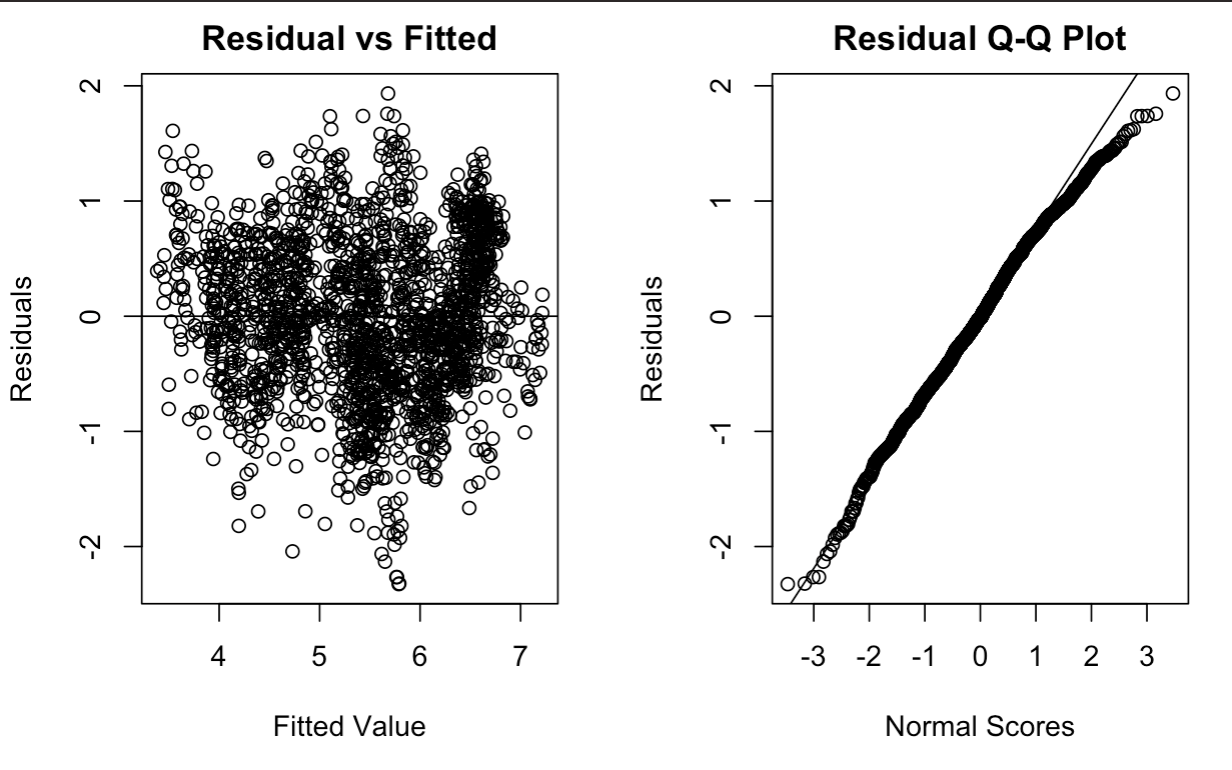


It can be seen from these plots that not only does economic stability have a correlation with happiness, it has an extremely strong positive linear correlation that has been proven over the 15 years with these scores. Following this visualization, it is made clear that potentially a simple linear regression model would be able to capture this trend and be able to predict relatively accurately the estimated happiness score of a country.



The black dots, representing each country’s GDP and score over the year, appear to follow along with the linear regression model shown by the red line. With very few outliers, there doesn’t seem to be any data points that are predicted wildly wrong. There is, however, one section in the top right of the graph where there is a strong concentration of data points above the model prediction. This can be explained by the concept that after a certain point, money is no longer able to buy more happiness and there must be other means to increase happiness. In addition, there appears to be countries that have high GDPs but low happiness scores. After further analysis, it is found that these data points mainly come from the country of Botswana. Botswana has a huge income inequality among its citizens which is most likely the biggest contributing reason as to why their GDP could be higher, but happiness lower.

In order to confirm that this model is statistically significant, tests and visualizations were performed. The following shows how the fitted values match up with the residuals.

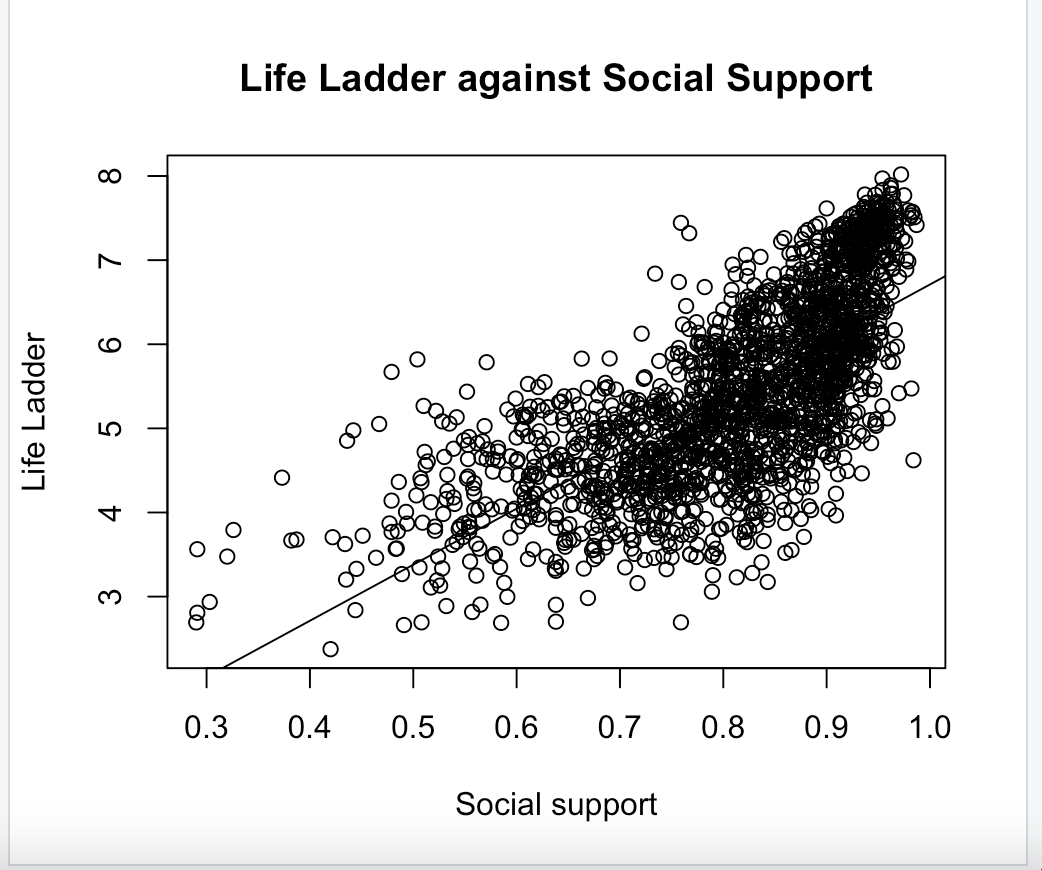


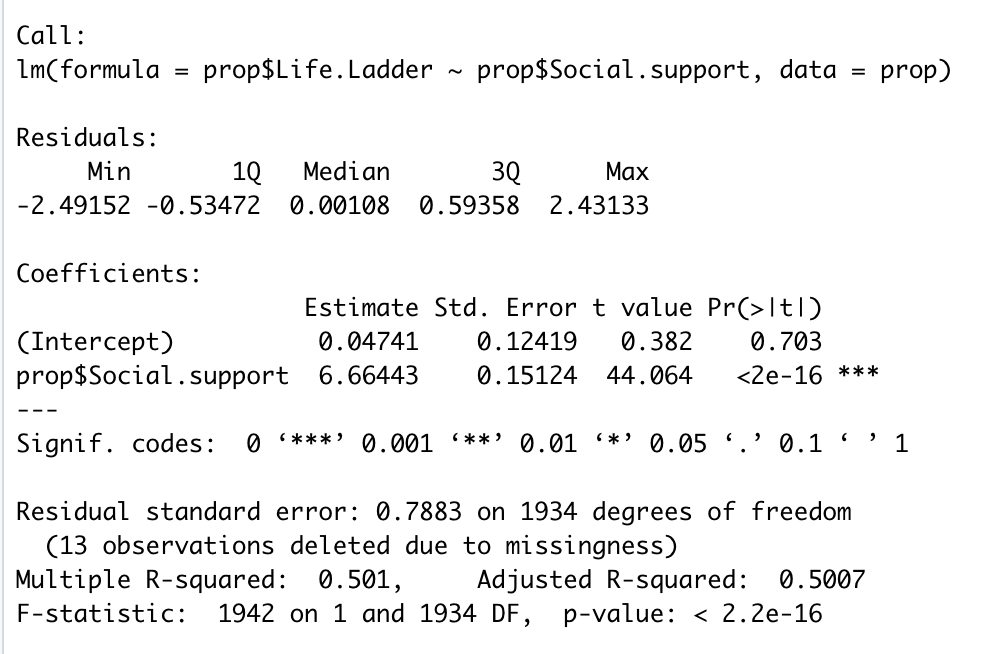
It can be seen from these plots that the model performs well on the data and is able to predict the happiness scores within 2 for almost every single country. The slight deviance at the top of the QQ plot can be explained by the outliers found previously in the linear regression plot. It’s important to not only consider one of these factors as there are many that could be playing a role in influencing the happiness level of a country.

* 1. **Social Support’s effect on Overall Happiness**

When considering the happiness of a country, the social support being received from those around the people in the country is definitely an attribute that should be factored in. From the surface, it would seem that happiness and social support should go together and have a positive correlation. Our simple linear regression analysis regarding the effect of social happiness on the happiness score of a country will allow us to support or oppose that notion.

Simple Linear Regression between Happiness Score and Social Support:

****

****

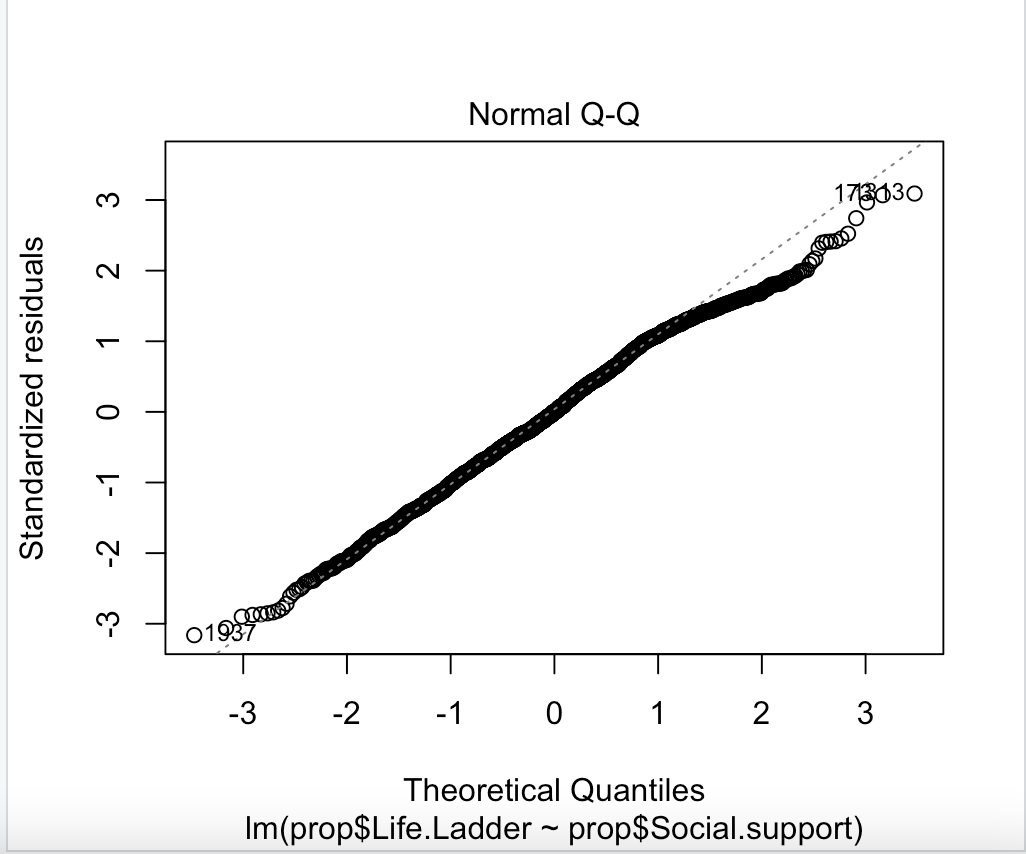
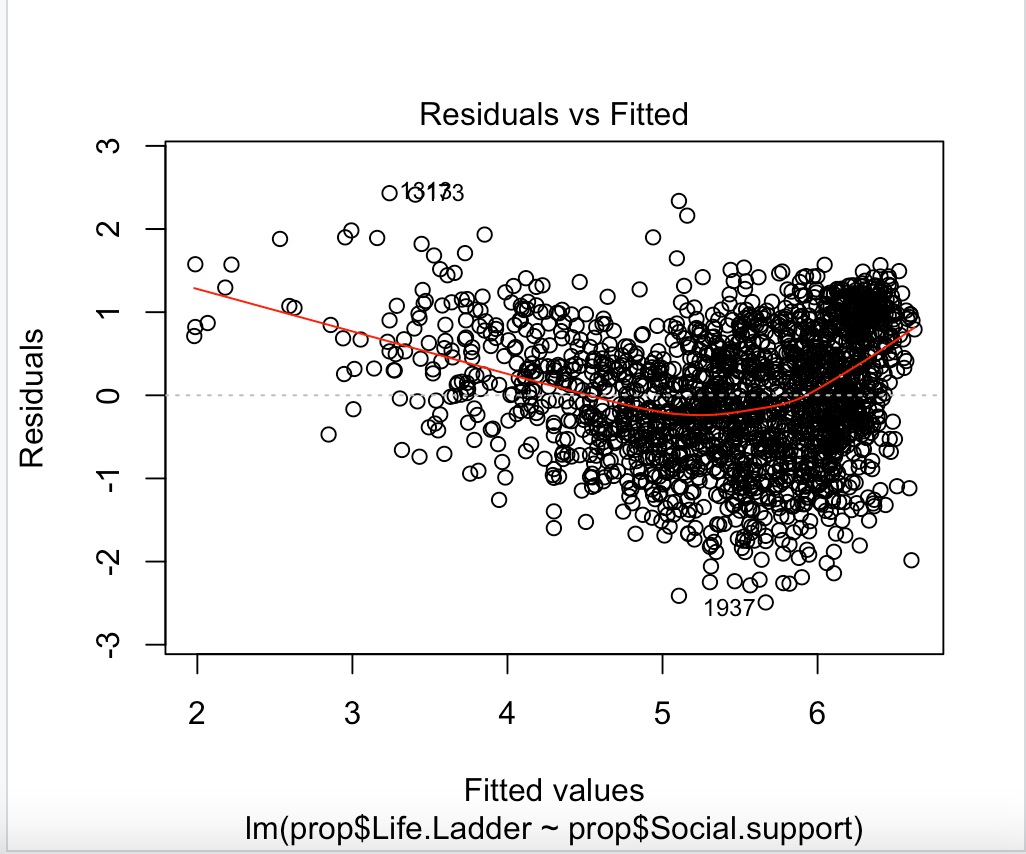
The R-square value for this simple linear regression model is 0.501, which implies that the predictor variable social support accounts for 50.1% of the variation in the outcome variable: the happiness score of a country. Since the R-square value is relatively large, it shows that the simple linear regression model can explain the response variable quite well.

The F-ratio for this model is 1942, with an extremely small p-value of less than 0.001. Therefore, we can conclude that the model with social support as the predictor variable yields better results compared to a model with the mean happiness score.

The output of the model shows the coefficient for the social support, which is 6.66. This means that the happiness score increases by 6.66 with every increase in social support. The intercept is 0.047. Hence, our model for predicting the happiness score with social support is:

Happiness Score = 0.047 + (6.66 \* Social Support)

After constructing the simple linear regression model, we have to check the assumptions we made to ensure they are valid.

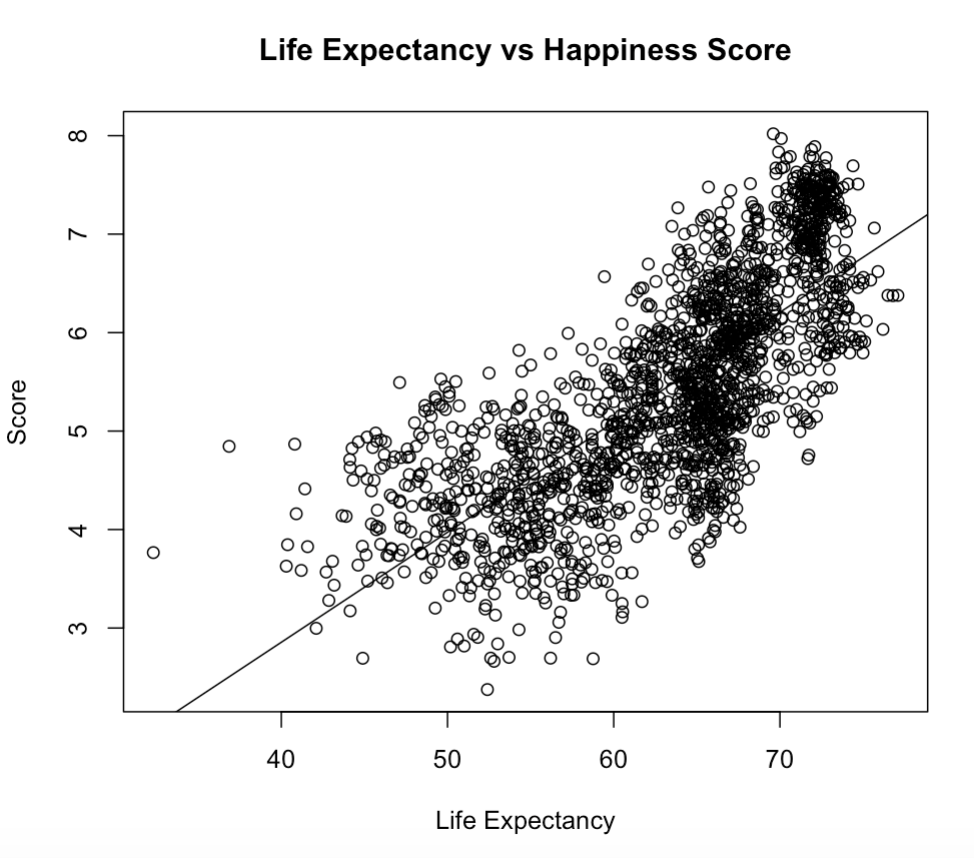


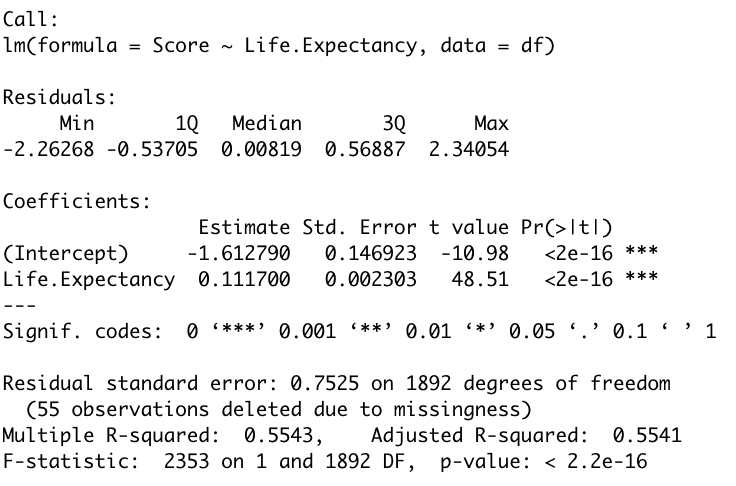
The left graph is a plot of the observed residuals against the fitted values. In this plot, the dots should be randomly located around the horizontal zero line, which is the majority of the case. However, there are some points at the top left that are considered outliers, which could be potentially explained by the fact that this model only considers one of the three significant attributes we are analyzing in this report, in relation to the happiness score.

The right graph is a normal Q-Q plot and the majority of the points lie on the y = x line, which shows that the distribution the points follow is fairly normal. On the other hand, the points on the top right of the graph slightly deviates from the dotted line. The outliers discussed in the residual plot might be the reason for this deviation.

* 1. **Life Expectancy’s effect on Overall Happiness**

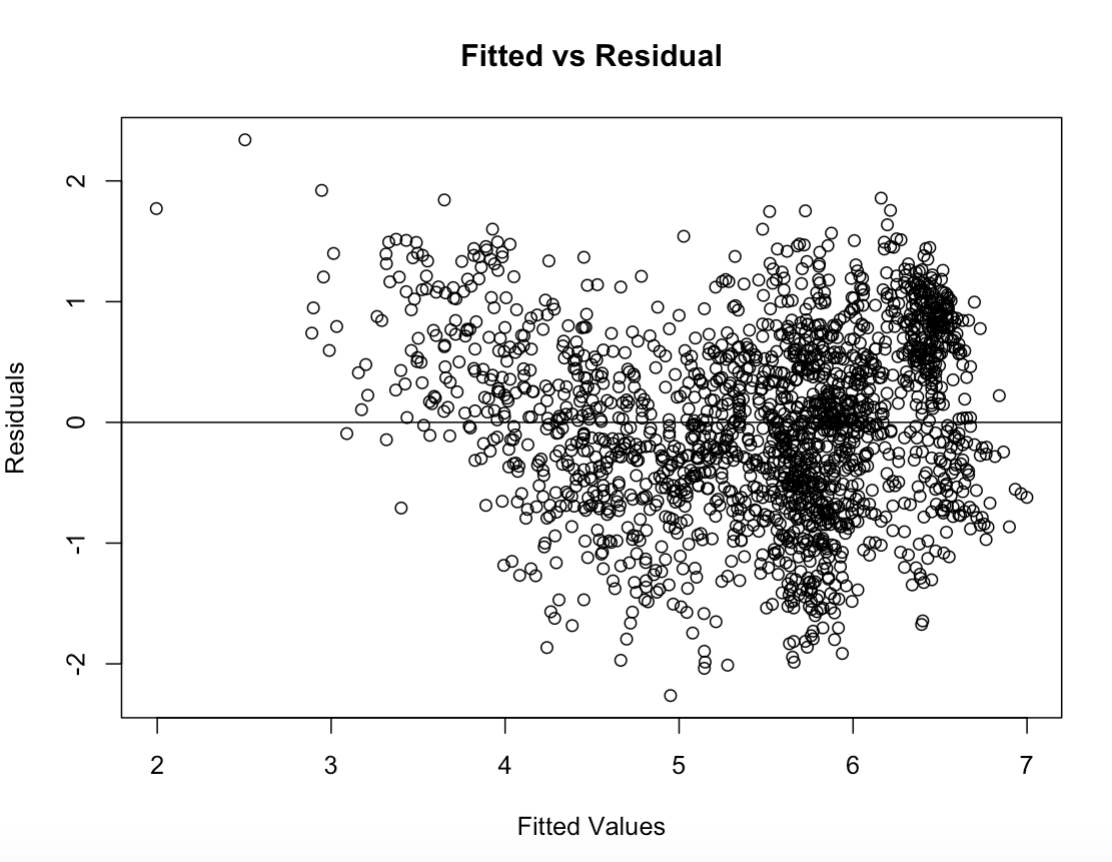
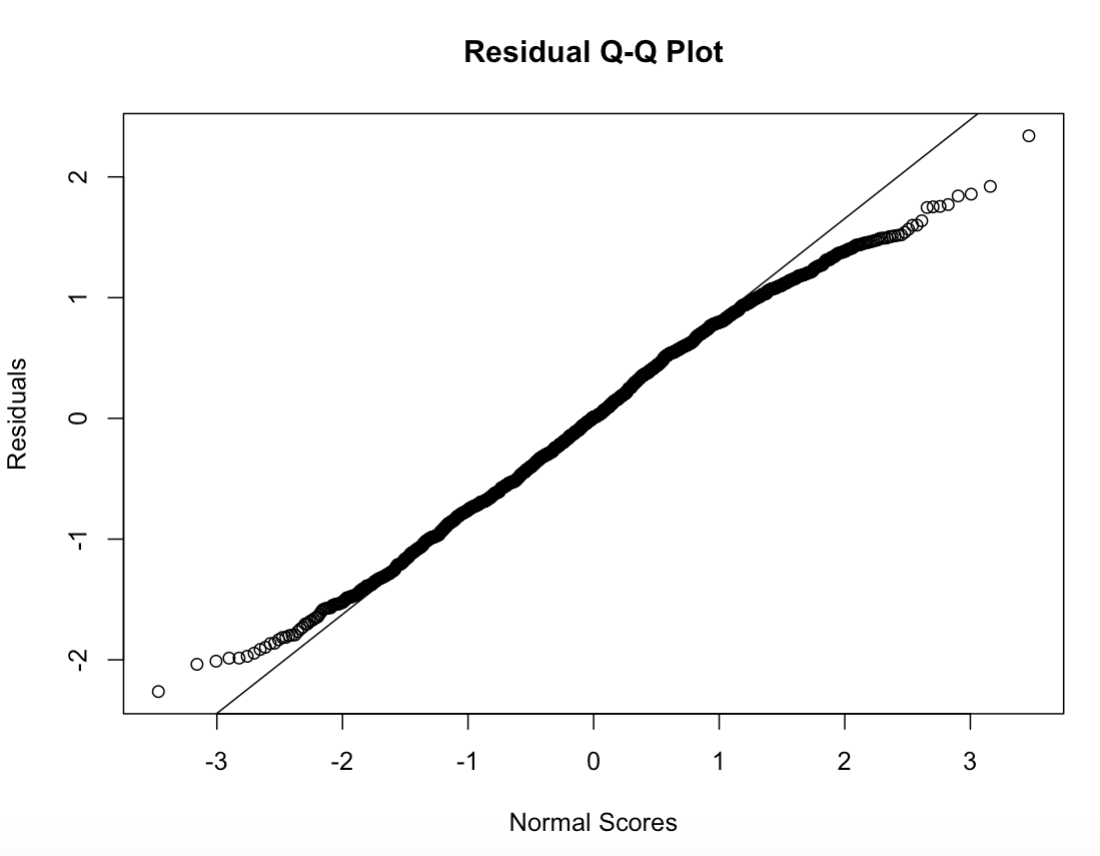
To analyze how the life expectancy of individuals within a country may weigh on that country’s overall happiness, we conducted a linear regression model. The graphical representation of the model is as follows:





Each circle represents the life expectancy and overall happiness score attributed to a specific country. As seen by the graph, life expectancy and happiness follow a linear distribution. Majority of the points hug the y=x line, with the exception of a dense portion in the top right. The R-squared value is 0.5541 which means that the regression function fits the data relatively well, but not perfectly. This could be explained by the fact that it’s rare for a given country to have an average life expectancy greater than 73. Thus, when you reach that point, life expectancy is no longer the only factor influencing overall happiness, resulting in a 1 point margin in happiness score, the 6.5-7.5 range. The F-statistic is 2353 and the p-value is much less than 0.01, thus our model using life expectancy to predict overall happiness is much better than simply using the average happiness score.

To try and confirm that this model is statistically significant we tested how the fitted values compare with the residuals. The visualizations are as follows:

The graph on the left represents the fitted values against the observed residuals. The plot contains points randomly scattered around the horizontal zero line, suggesting that, for the most part, the assumption that the relationship is linear is reasonable. A couple points in the top left corner stand out from the pack suggesting that there are a few outliers, as seen in the above linear regression plot. This can be justified by the fact that life expectancy isn’t the only factor affecting overall happiness.

The graph on the right represents the Q-Q plot which portrays the majority of the points following the y=x, representing that the data is fairly normally distributed, except for the tailends which make up the same set of outliers explained above. From these plots we can conclude that the model is able to predict the happiness score within 2 adequately for all of the countries.

After determining that life expectancy and score follow a linear relationship, we decided to graph the regional indicators against the average life expectancy of all of the countries found within that region in order to find the happiest regions.

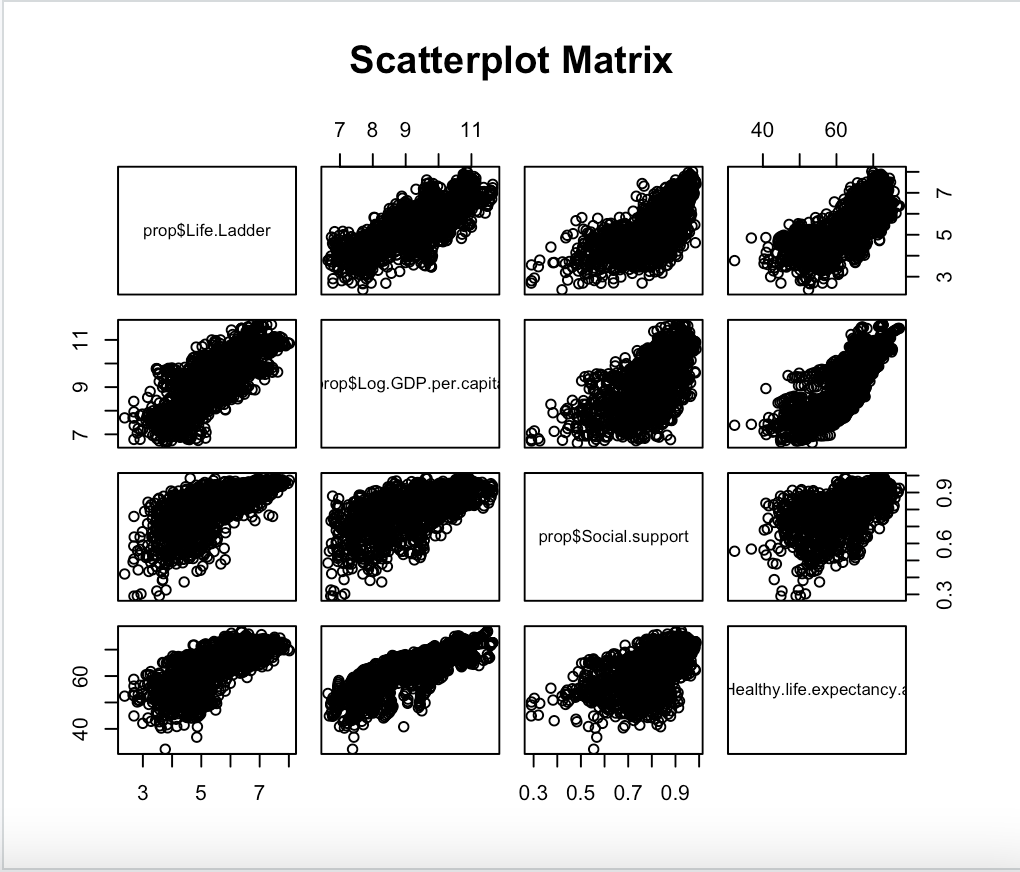


In this plot you can see that the largest average life expectancies can be found in Central and Eastern Europe, East Asia, North America and ANZ, and Western Europe. Thus we concluded that the happiest countries, in regards to life expectancy, can be found in those regions. These results are possibly representative of the better health care systems, reduced violence and drug and/or alcohol use, as well as the cheap, or free, education found in these regions.

1. **Multi Linear Regression Model**

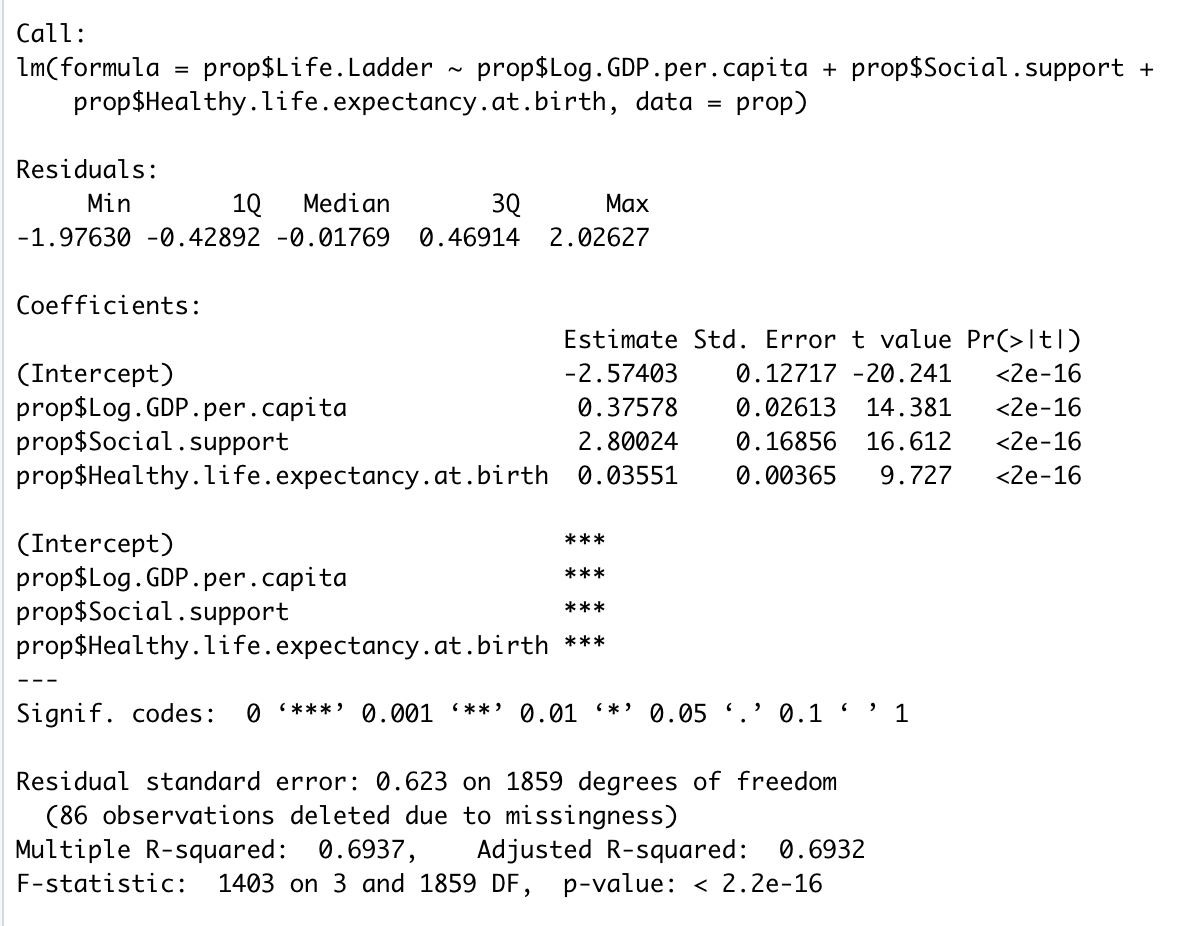
Analyzing something using only one attribute at a time while there are various factors contributing to it is not the most effective method. Thus, after conducting simple linear regression analysis on each of the three attributes, we created a multiple regression model that integrates all three attributes to give us a better sense of the role each attribute plays in the happiness score of a country. This model includes the outcome variable, happiness score, plotted against three predictor variables: economic stability (GDP), social support and life expectancy.

We created a scatterplot matrix to visualize the relationship between the variables.



In general, there is a positive linear correlation between the three predictor variables and the outcome variable. The scatter plot matrix allows us to compare not only the relationship of each predictor variable to the outcome variable, but also the predictor variables with one another. Being able to conduct the comparison of these relationships side by side in one diagram makes it easier to draw conclusions.

Our multiple linear regression model with economic stability, social support and life expectancy as predictor variables, and happiness score as the outcome variable is shown below.



The R-squared value for the model is 69.37%, which is considered fairly good. One of the assumptions of multiple linear regression is that the predictor variables are numeric, which is consistent with the data we are utilizing. An R-squared of 0.6937 implies that the economic stability, social support and life expectancy is responsible for 69.37% of the happiness score variance.

The output shows the coefficients for economic stability, social support and life expectancy are 0.376, 2.80 and 0.0355 respectively. The intercept is at -2.57. The equation for the multiple linear regression model predicting the happiness score using the three attributes mentioned is:

Happiness Score = -2.57 + (0.376 \* GDP) + (2.80 \* Social Support) + (0.0355 \* Life Expectancy)

**5.1 Outliers and Influential Cases**

In order to find outliers we looked at the standardized residuals. Under the assumption that standardized residuals greater than 3 are deemed outliers, we took the percentage of values with standardized residuals outside of this range. Obtaining that 0.3% of outliers are not significantly affecting our model.

Next, we calculated the cook’s distance of the values. No values have a cook’s distance greater than 1, shown by the maximum and minimum values below. This indicates our data has no influential cases.

|  | Max | Min |
| --- | --- | --- |
| Cook’s distance | 1.000e-9 | 2.212e-2 |

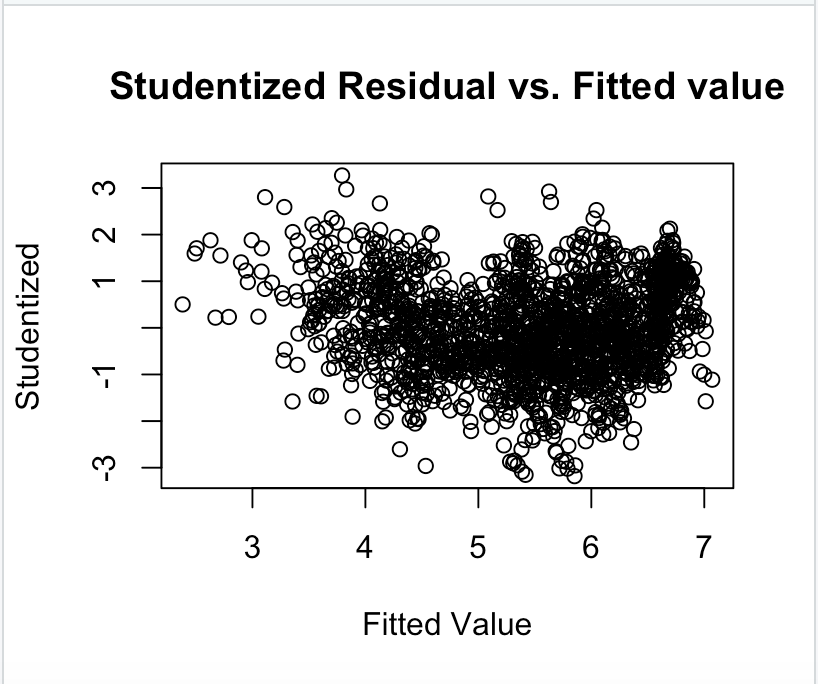
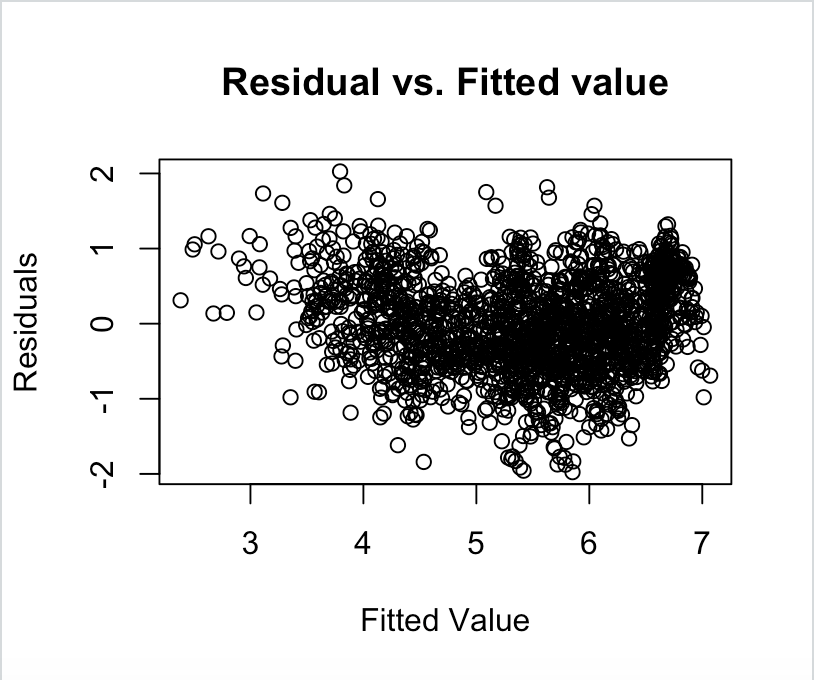
**5.2 Multicollinearity**

Looking at the effects of the Multicollinearity assumption, we calculated the variance influence factor (VIF) of our three significant variables. VIFs between five and ten are considered significant with poorly estimated regression coefficients. We obtained the values below.

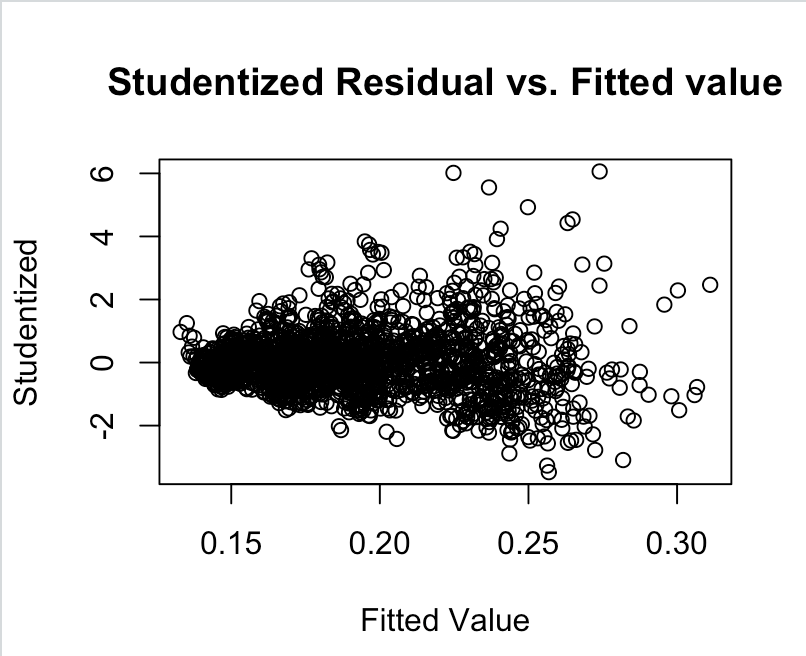
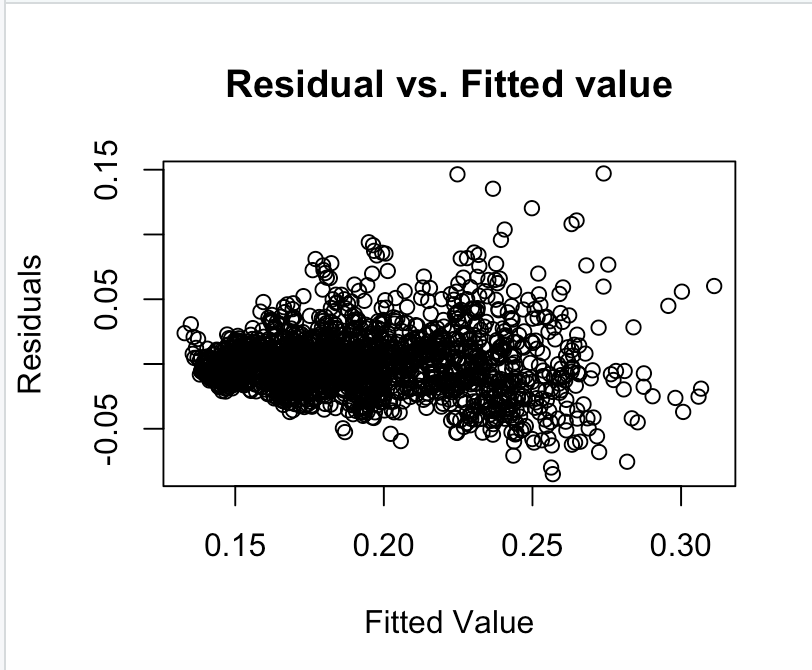
|  | GDP | Social Support | Healthy life expectancy |
| --- | --- | --- | --- |
| VIF | 4.36 | 1.94 | 3.60 |

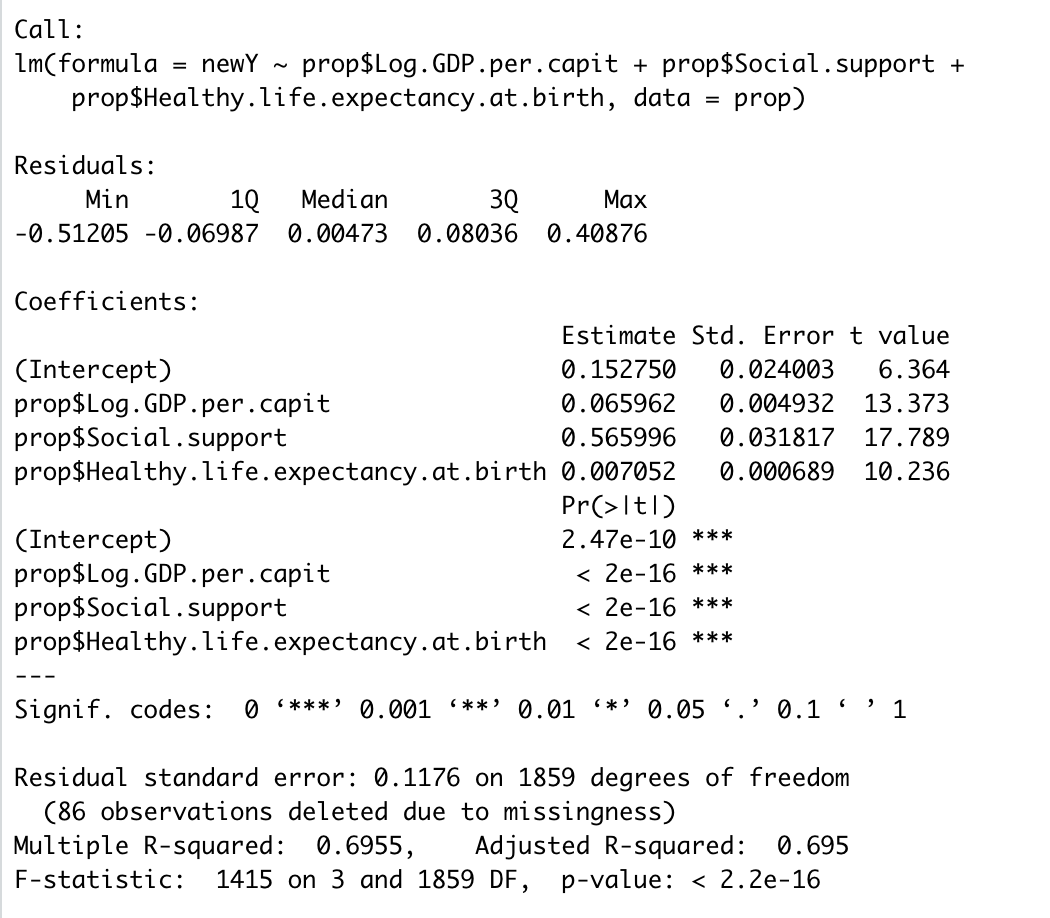
Because none are greater than five or ten, there is no multicollinearity in our data.

**5.3 Residuals**



Shown in both graphs, the residuals vs. the fitted values show somewhat of a U-shaped pattern. In order to fix this problem of non-linearity, we analyzed the transformation of y to the natural log of y on the model.





Both plots show more randomized patterns around the horizontal zero line, indicating this transformation improved our model. The r-squared value of our new model is 69.55% which is higher than the previous models r-squared value, 69.37%, proving there is some non-linearity in our dataset.

1. **Conclusion**

Based off of our analysis, we were able to determine that:

1. GDP, Social Support, and Life Expectancy all have a positive, linear relationship with the overall happiness found in a country.
2. These 3 variables are the most correlated with the score so you should focus most of your attention on them when deciding where to move, implementing state/national policy, etc...
3. Each one provides good insight into predicting how happy you will be, yet there are multiple variables that go into the score.
4. Thus, someone who is planning on moving to a country should look at all of the factors before making a decision, but place the largest weight on GDP, social support, and life expectancy during the process.