Estimations of Happiness by Country

Leon Weingartner

Introduction

The measurement of happiness is a difficult and ambiguous task; however, this analysis sets out to quantify, from an objective standpoint, what indicators can be used to approximate a country's level of happiness and where it ranks relative to others. National governments can use these results to better understand why some countries are happier than others. Improving the wellbeing of humanity is a tricky and complex problem to solve but narrowing our search down to just a few key variables allows us to estimate a numerical value that describes the average happiness score of an entire country. The relevance of this study becomes apparent when considering changes in governance. After all, government is an established framework with the purpose of allowing people to live, in many cases, a free and fulfilling life. one could consider many factors that would contribute to one's happiness. However, we will focus on a few variables pertaining to GDP, social support, life expectancy, and how government is perceived by the general public. These indicators are all dependent on the way any particular country is governed. Understanding the connection between these variables and happiness can inform government policy in the future.

Study Focus

This analysis will incorporate data from 156 countries. The comprehensive collection of data being used was compiled by the World Happiness Report which has been accumulating data every year since 2012. The rankings of all 156 countries, based on happiness scores, are given where the Scandinavian countries seem to perform very well every year. For this paper we will research the data from 2018 as it integrates more factors than some years and is not heavily affected by the ongoing pandemic from COVID-19. There is of course room for a more detailed interpretation regarding happiness such as the differentiation of city and rural living, engagement in local communities, willingness to volunteer, addiction to certain substances, immigrants and citizens, religious affiliations, environmental pollution, technological advancements, unemployment rate, and so much more. In terms of specificity I'm interested in studying the ranks of happiness by geographical area. Are certain continents significantly happier than others? Since there is more than one factor contributing to a country's well-being, this study will contain a model to predict happiness by using a multiple linear regression instead of just a simple linear regression.

Data

This dataset was imported from Kaggle containing information for 156 countries in 2018. The table below showcases the following variables used in this study:

Variable	Name	Description
Overall Rank	Overall.rank	Country's happiness score rank
Country or Region	Country.or.region	Country name
Score	Score	Happiness score
GDP per Capita	GDP.per.capita	Country's GDP (Gross Domestic Product) per capita
Social Support	Social.support	National average of binary responses regarding social support
Healthy Life Expectancy	Healthy.life.expectancy	Ratio of healthy life expectancy and life expectancy
Freedom to Make Life Choices	Freedom.to.make.life.choices	National average of binary responses regarding freedom
Generosity	Generosity	Measures generosity by donations
Perceptions of Corruption	Perceptions.of.corruption	National average of binary responses regarding corruption in government and businesses

Origin of Data

Happiness Score:

This score is evaluated by taking the overall average answer to the famous Cantril ladder question which assesses an individual's quality of their current life on a scale from 0 to 10. Most countries that were polled have an annual sample size of about 1000 participants where the average score is taken from the cumulative answers from at most 3 subsequent years (typically 2000 to 3000 samples per country).

GDP per Capita:

This data was taken from the World Development Indicators released by the World Bank. The previous year's GDP numbers had not been released in time, so a forecast of real GDP growth was used from the World Bank's Global Economic Prospects after adjusting for population growth. The actual data used followed the equation c*ln(GDP per capita) for some constant c as this fits the data substantially better than just GDP per capita.

Social Support:

This value was obtained from the Gallup World Poll as a national average to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" (binary response 0 or 1).

Healthy Life Expectancy:

This data was gathered from the World Health Organization (WHO) in 2012. The values used in this data set are the ratios of healthy life expectancy to life expectancy.

Freedom to Make Life Choices:

This value was obtained from the Gallup World Poll as a national average to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" (binary response 0 or 1).

Generosity:

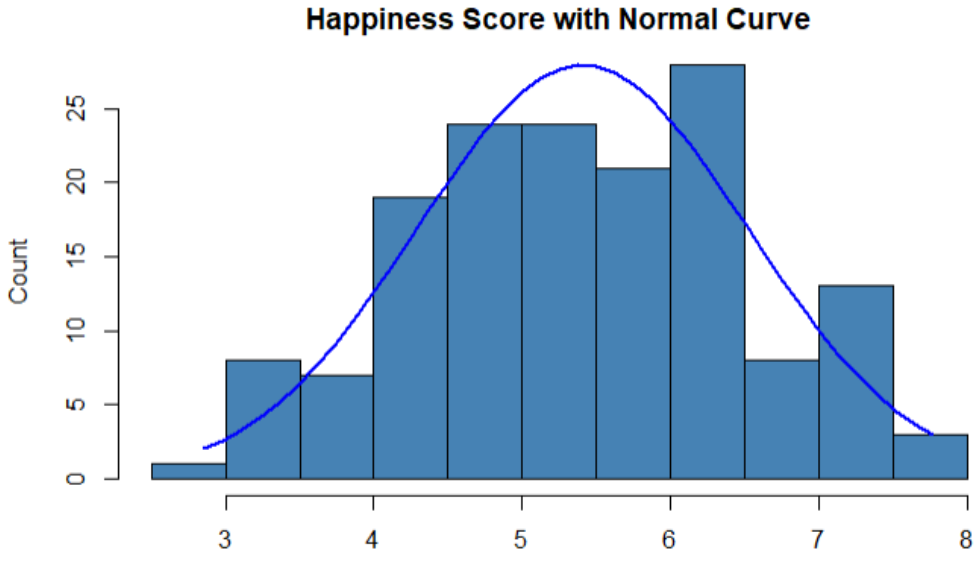
This value was obtained from the Gallup World Poll as the residuals after regressing the national average value to the question "Have you donated money to a charity in the past month?" with GDP per capita.

Perceptions of Corruption:
This value was obtained from the Gallup World Poll as a national average to the two questions “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?” (binary responses 0 or 1).

Analysis and Testing

Our goal is to build a statistical model that will allow us to predict the level of happiness in a country based on a few impactful indicators that we will research throughout this section. After determining the factors, we will use in a multi linear regression model, we can begin to test this and predict future values. The model we end up using should not be prone to overfitting as our sample size is substantial and we only have a short list of variables to play with.

First lets better understand how our distribution of happiness looks on the scale of 0 to 10. I’ve broken each interval with a length of ½ and tallied the scores that fall into each bound. The graph below shows the distribution with the normal curve having matching means and standard deviations for comparison. It’s clear that our sample size is not big enough to closely follow the normal curve however there is an obvious bell shape to it, which is expected.

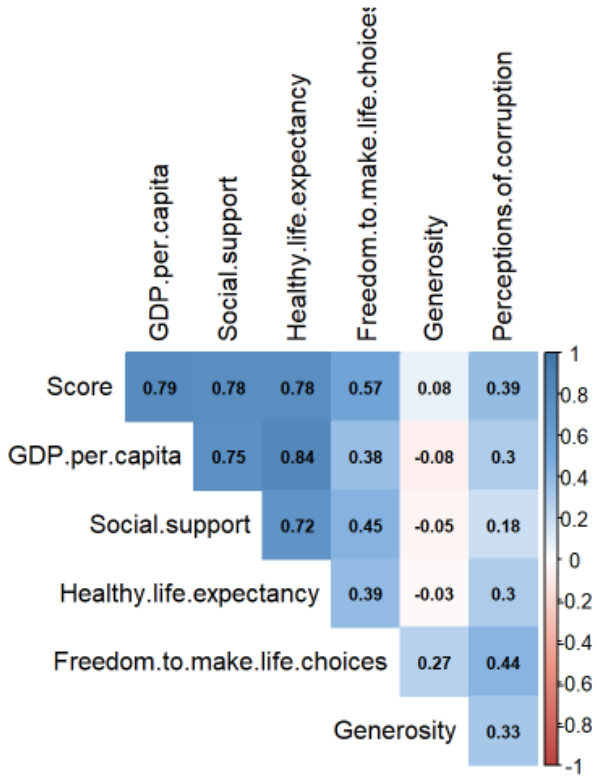



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.853   4.545   5.380   5.407   6.184   7.769
```

We see that the median and mean very close together suggesting very little skewness. In the upcoming multilinear regression model, we will not need to use the log function to normalize the happiness score.

Correlation

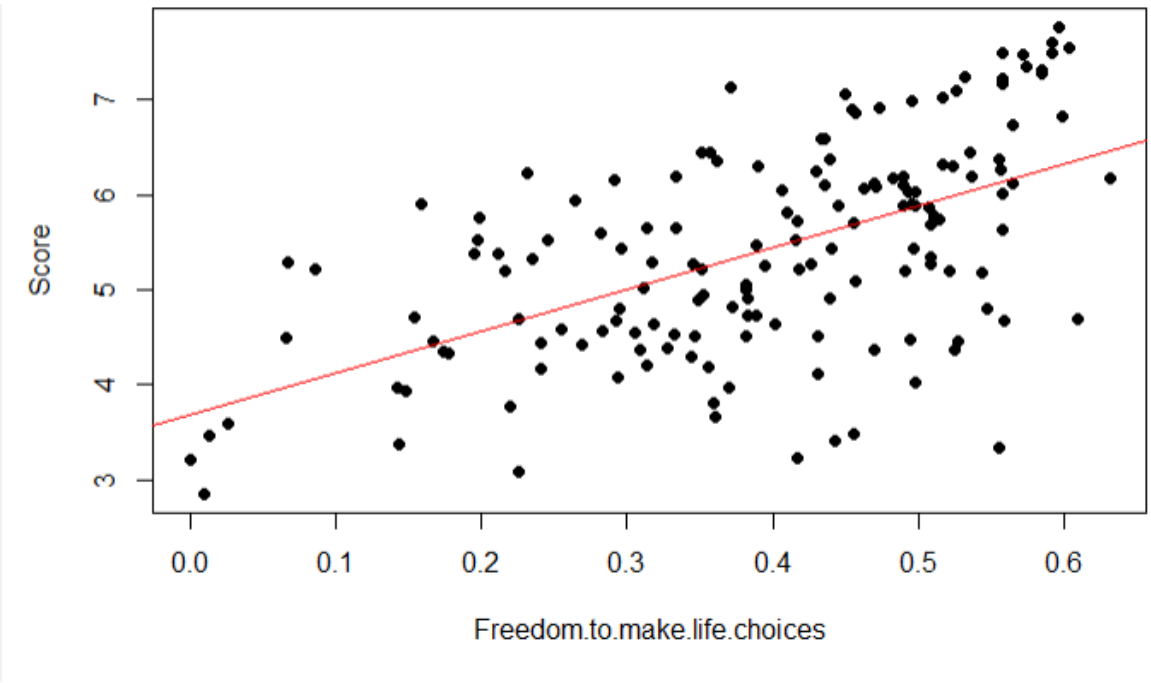
We can use the correlation coefficient to see which variables are strongly related to happiness. Using this coefficient, denoted by r, gives us the linear relationship between 2 sets of data where r is in the interval [-1,0) indicates a negative relationship, r in the interval (0,1] indicates a positive relationship, and 0 indicates no relationship. Below shows a graph of correlation coefficients color coded by the given r values along with the legend on the right-hand side. This shows the relationship between any 2 variables which gives us a preliminary idea of which factors we will use in our upcoming model.



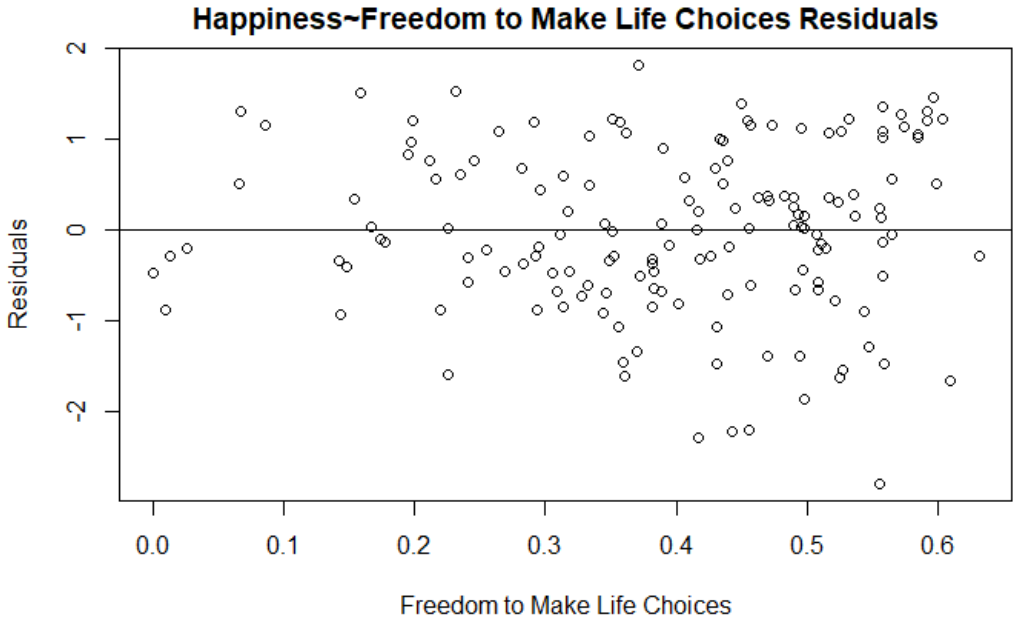
Our coefficients of interest lie on the first row containing the relationships between the happiness score and every other variable in the data set. Interestingly, GDP per capita, social support, and healthy life expectancy seem to have the strongest correlation against happiness. Whereas generosity and perceptions of corruption have little to no correlation. The only variable we could expect to have a negative correlation would be with perception of corruption, however neither generosity nor perceptions of corruption seem useful in predicting a happiness score. We will be taking a more in depth look at the variables GDP.per.capita, Social.support, Healthey.life.expectancy, and Freedom.to.make.life.choices before we develop our model.

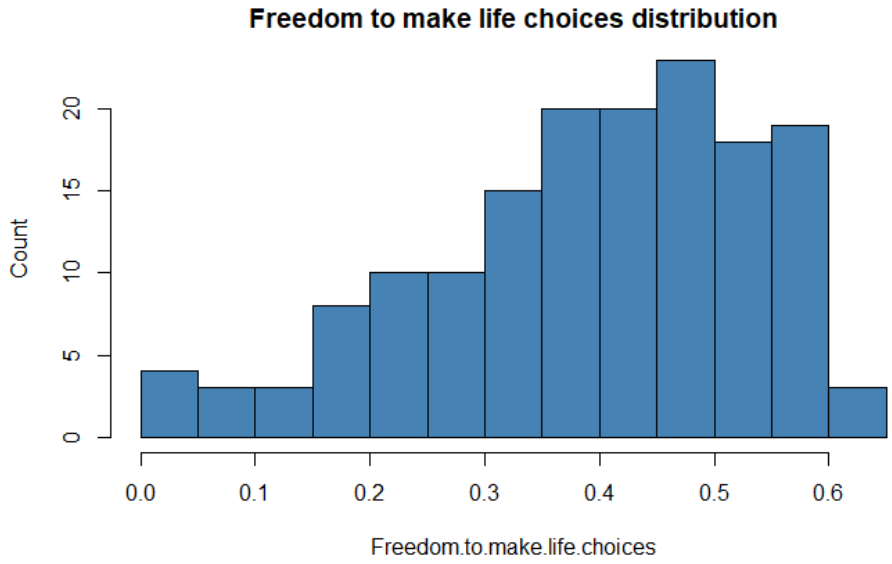
Freedom to Make Life Choices

With a correlation coefficient r = 0.57, our coefficient of determination is calculated by squaring this value to get 0.3249. That is to say roughly 32% of the total variation of happiness in this dataset can be explained by the freedom to make life choices surveyed with a binary response as explained above under the origin of data section. Although it's our lowest r^2 value we will be using to determine our model, 32% is still a significant part to include when trying to predict happiness. Below depicts a scatterplot of our average freedom to make life choices value (x-axis) against happiness score (y-axis) for all 156 countries. I've also included a simple linear regression that allows us to visualize our coefficient of determination.



We can see here that our relationship is somewhat linear however our values are not closely in line with our simple linear regression. This means our ability to predict happiness from freedom to make life choices alone is quite difficult and this is even more apparent when plotting the residuals of this regression. We see below that almost all of our values lie within a distance of 2 (absolute value). Since our score is on a scale from 0-10, the possible error when only using this variable to predict happiness is too large.



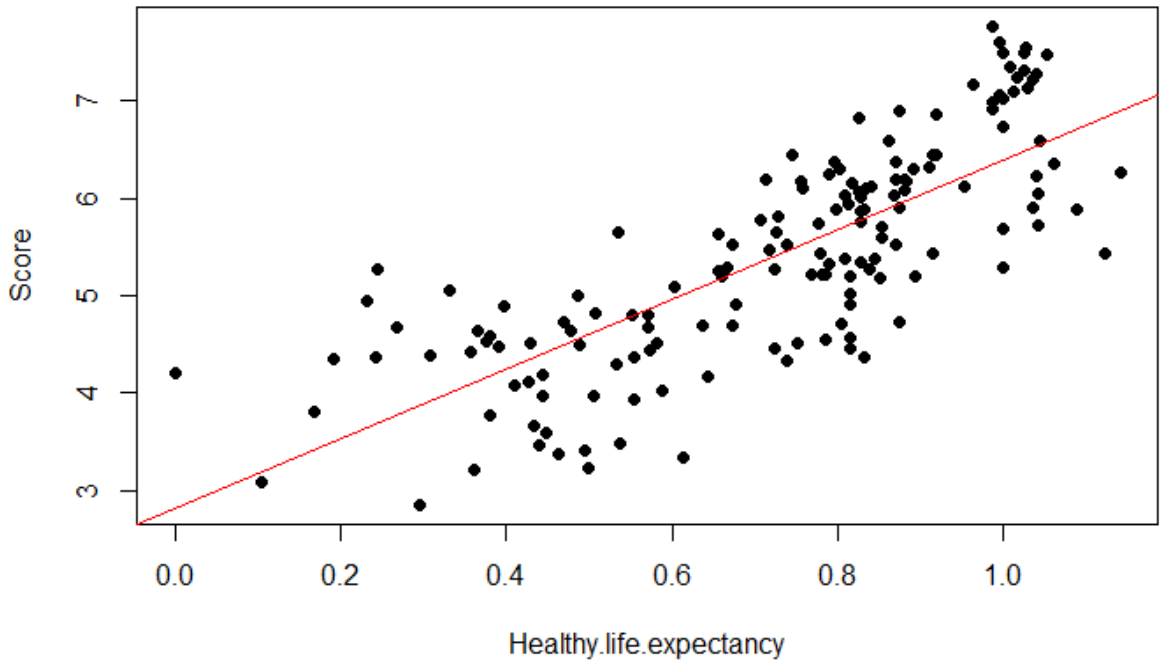



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.3080  0.4170  0.3926  0.5072  0.6310
```

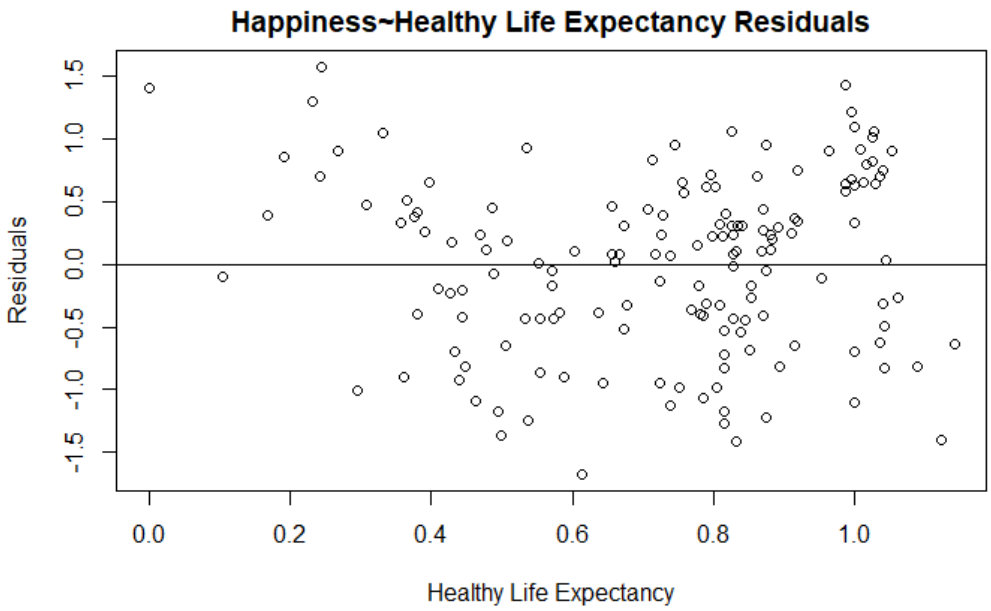
The distribution of values for freedom is certainly skewed to the left. With possible values from 0 to 1 resulting from a binary response question, it seems that in the majority of countries, generally speaking, people are not satisfied with their freedom to make their own life choices.

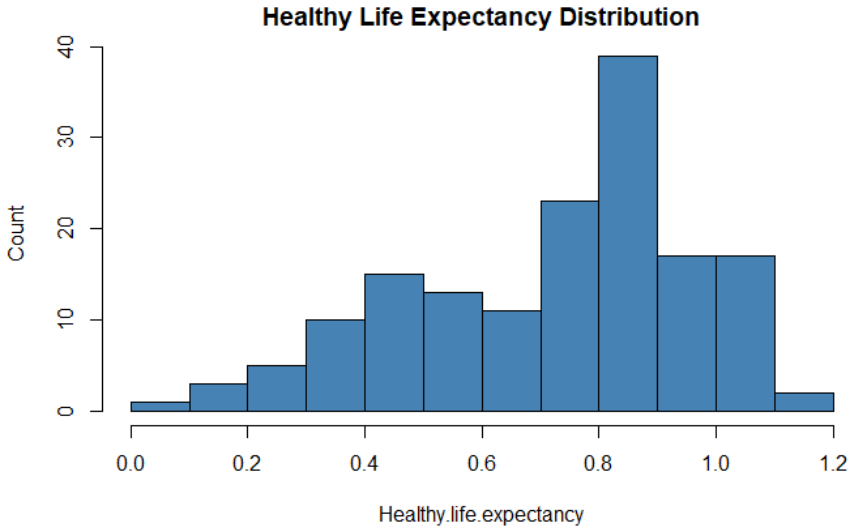
Healthy Life Expectancy

With a correlation coefficient r = 0.78, we get a coefficient of determination of 0.6084. This is easily a significant relationship (p-val of 0.00091) and worthy of investigating. Below depicts a scatterplot of our average healthy life expectancies (x-axis) against happiness scores (y-axis) for all 156 countries. Similarly, as with our last variable, I've included a simple linear regression that allows us to visualize our coefficient of determination.



With this graph its evident that healthy life expectancy follows a much stronger linear relationship with happiness than freedom. The residuals are conspicuously closer to our simple regression line and incorporating a variable such as this with our multiple linear regression model would definitely show to be an improvement. Below shows the residual graph for healthy life expectancy and our room for error when following the simple regression line. Almost all of our values lie within a distance of 1.5 and is seemingly a better choice when using just 1 variable to predict happiness.

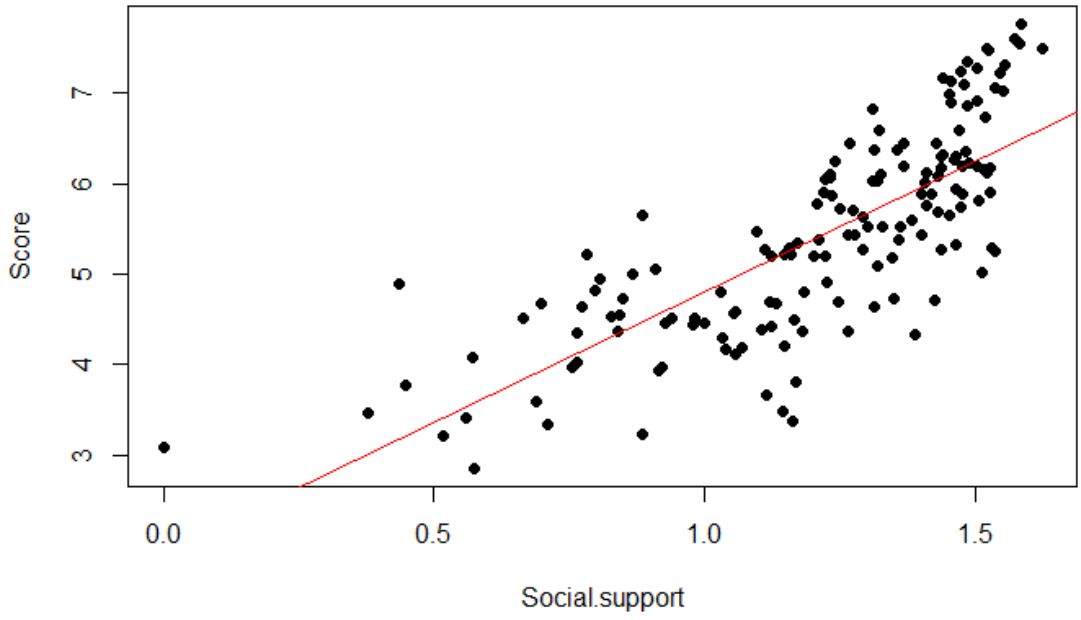




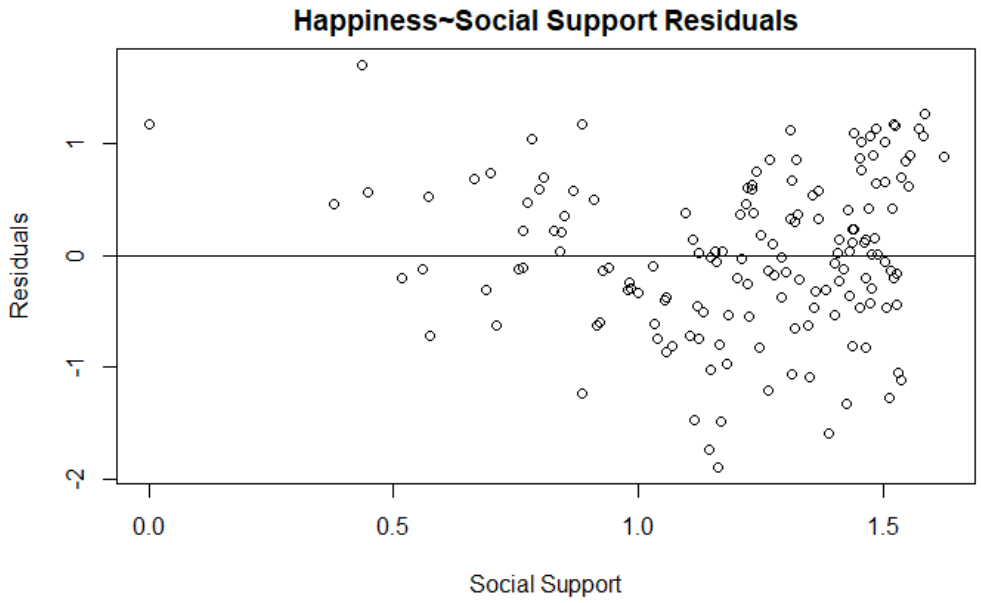

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.5477  0.7890  0.7252  0.8818  1.1410
```

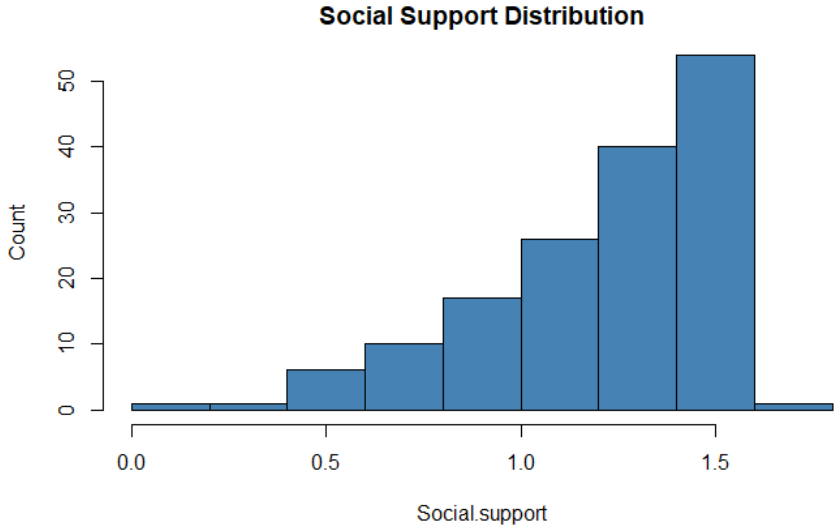
Social Support

We find that our correlation coefficient is the same as healthy life expectancy with a value of r = 0.78, thus we have a coefficient of determination of 0.6084. As mentioned above, these data points are a result of another binary response question relating one's relationship with family and friends. The choice for these survey questions may seem somewhat arbitrary however, the responses seem to show a very significant relationship with happiness. The scatterplot below depicts values averaged for social support (x-axis) against happiness scores (y-axis). Following the same evaluation style for each variable, I've included a simple linear regression, residual graph, and distribution chart for social support to help us better understand what indicators we can use to predict happiness.



Notice our distribution of social support is heavily skewed to the left indicating how countries respond when asked about family and friends. With a larger cluster of data following the higher values of social support, one could speculate the importance of trust within the individualized domain of social interactions to whom we are most comfortable with. Along with this expected outcome we find a quite interesting relationship with our recorded happiness scores. The residual graph shows the majority of our data lying within a distance of 1 from the simple linear regression which further showcases the promising potential of using social support in our multilinear regression model.

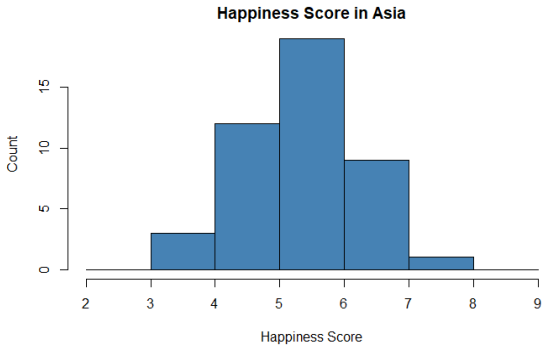


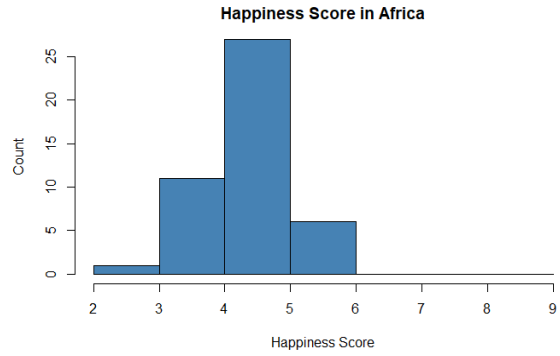


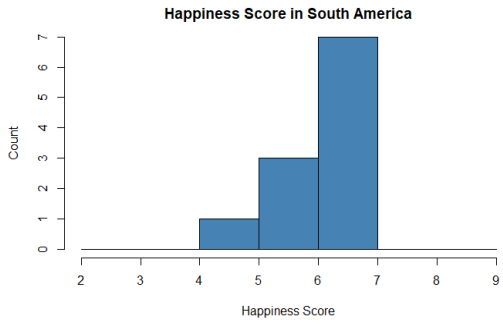

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   1.056   1.272   1.209   1.452   1.624
```

Further analysis on the remaining variable GDP per capita shows very similar results from the previous 2 variables that were researched as it has a similar correlation coefficient and is outlined in the origin of data section from page 2. Before constructing the multilinear regression model, I wanted to explore the happiness distributions separated by continent to see if there are any significant differences and possibly what research can come out of it.

Happiness Score Distribution by Continent








```
[1] "Europe"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.332   5.603   6.149   6.268   7.021   7.769
[1] "Asia"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.203   4.699   5.254   5.265   5.888   7.139
[1] "Africa"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.853   3.975   4.461   4.368   4.722   5.888
[1] "North America"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.597   5.883   6.287   6.152   6.669   7.278
[1] "South America"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.707   5.761   6.086   5.945   6.242   6.444
```

The above graphs show the distribution of happiness scores separated by continent. Europe clearly shows a lead in happiness with a mean of 6.268 and a minimum of only 4.332. Africa on the other hand is clearly behind the rest of the world with only a mean of 4.368 and a minimum of 2.853. This discrepancy is immediately obvious and is worth researching more about. North America and South America also show promising results, however, the sample sizes for these continents are not large enough to compare in a useful manner. After exploring factors that have a high correlation with happiness, it becomes transpicuous that Africa, most notably having a lower GDP per capita and healthy life expectancy, ranks poorly with respect to other continents.

Multilinear Regression Model

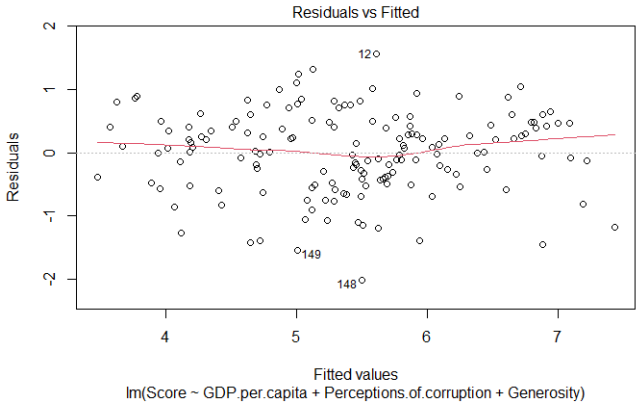
After our preliminary research on several useful variables, we will be using the 4 factors to construct a model to help us predict happiness consisting of GDP per capita, social support, freedom to make choices, and healthy life expectancy. We will use several combinations of these variables, as well as some others, to showcase which indicators produce the most simplistic model while retaining high predicting power when attempting to estimate happiness for a given country. Multilinear regression models follow the general equation

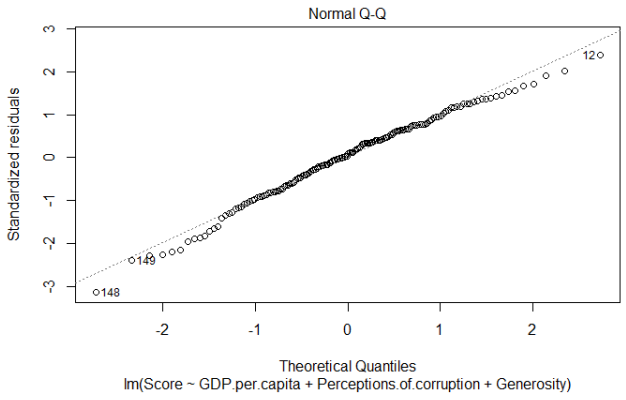
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

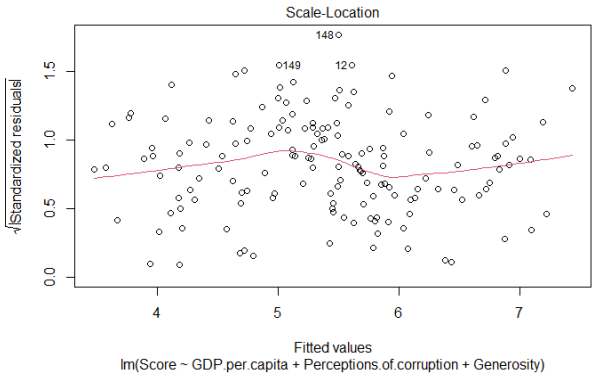
such that k describes the number of predicting variables and ϵ is an error term which we can assume to be normally distributed with a mean of 0 and a constant variance.

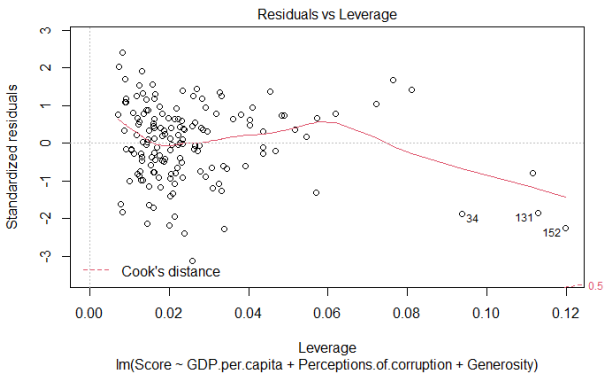
```
Call:
lm(formula = Score ~ GDP.per.capita + Perceptions.of.corruption +
    Generosity, data = happinessData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01491 -0.42212  0.04152  0.44160  1.55664

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.1050     0.1722  18.031   <2e-16 ***
GDP.per.capita              2.1341     0.1408  15.160   <2e-16 ***
Perceptions.of.corruption   1.4850     0.6256   2.374   0.0189 *
Generosity                  1.1158     0.5944   1.877   0.0624 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6533 on 152 degrees of freedom
Multiple R-squared:  0.6622,    Adjusted R-squared:  0.6556
F-statistic: 99.34 on 3 and 152 DF,  p-value: < 2.2e-16
```

Above shows the summary of this particular multilinear regression model. Both perceptions of corruption and generosity have high p-values and suggest this is not a very strong model following an R-squared value of just 0.6622. This model was experimented with merely to demonstrate the importance of picking highly correlated variables to represent a sufficient multilinear regression model. Below depicts the residual and qq plots.









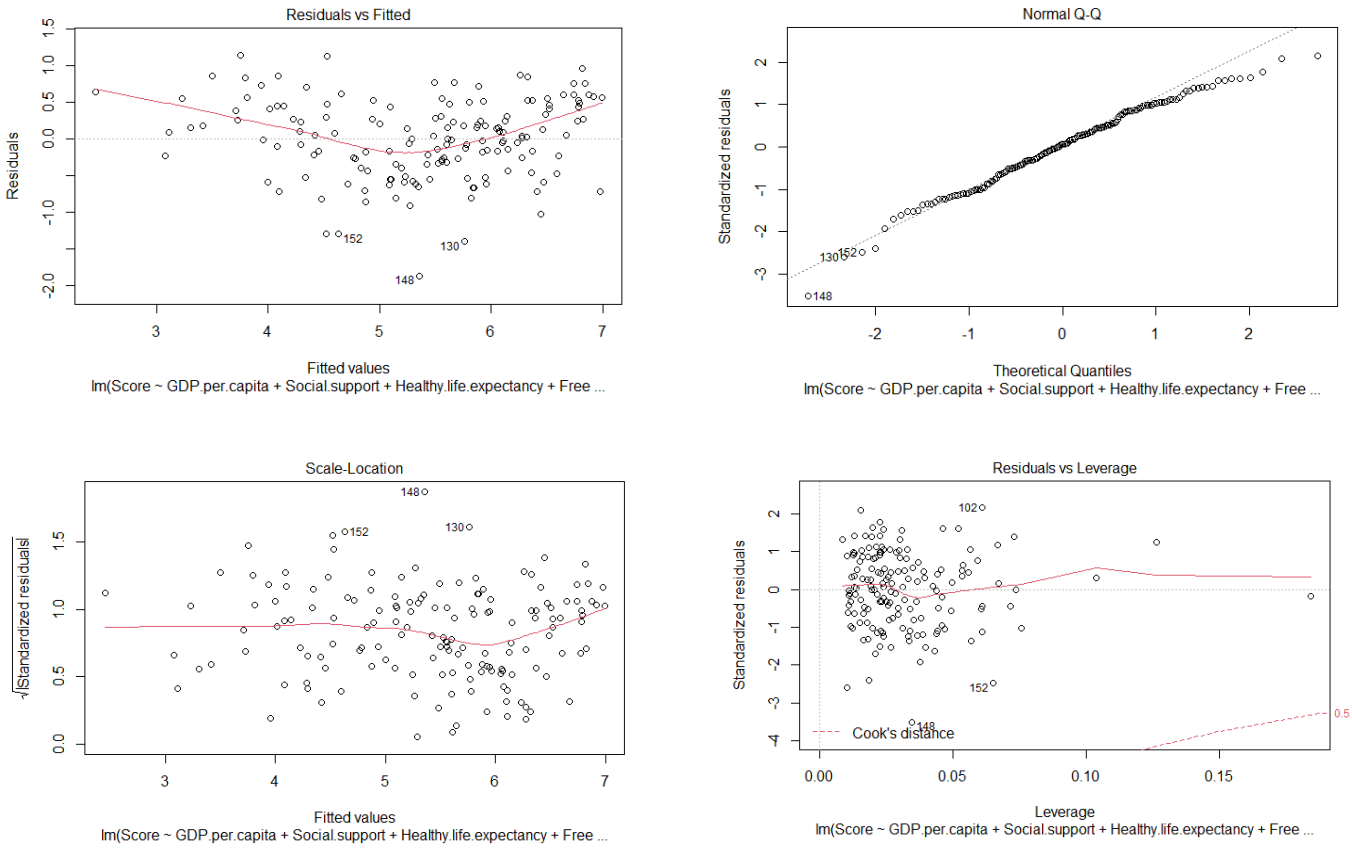
Although these plots show good signs of using a linear fit, this model can be greatly improved upon just by switching out certain variables. Signs for a good multilinear regression model include residuals following a normal distribution, an even scatter of data under the residuals vs. fitted plot indicating a constant variance, and a relatively flat trend when plotting for scale-location. A significant problem is prominent in the residuals vs. leverage plot where we see a number of outliers that influence the weights of our predictors. For these reasons, an alteration within our model can satisfy all of the assumptions for a multilinear regression model. As originally planned, we will use the four variables discussed and analyzed above to give us the equation

$$Score = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Where we have the following representations: x_1~GPD per capita, x_2~Social support, x_3~Healthy life expectancy, and x_4~Freedom to make life choices. The summary for this model is shown below along with the residual plots.

```
Call:
lm(formula = Score ~ GDP.per.capita + Social.support + Healthy.life.expectancy +
    Freedom.to.make.life.choices, data = happinessData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86584 -0.34594  0.03403  0.43676  1.13076

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.8921     0.1994   9.491  < 2e-16 ***
GDP.per.capita                 0.8105     0.2165   3.745 0.000256 ***
Social.support                 1.0166     0.2347   4.331 2.70e-05 ***
Healthy.life.expectancy        1.1414     0.3373   3.384 0.000910 ***
Freedom.to.make.life.choices   1.8458     0.3404   5.423 2.28e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5398 on 151 degrees of freedom
Multiple R-squared:  0.7709,    Adjusted R-squared:  0.7649
F-statistic:   127 on 4 and 151 DF,  p-value: < 2.2e-16
```

With an R-squared value of 0.7709 we can interpret this model to explain roughly 77% of the variation of happiness scores in our dataset. To compliment this strong relationship, we have p-values very close to 0 suggesting that these predictor variables are statistically significant.



In the residuals vs. fitted plot, our data points lie closer to our regression which is indicative of our increased R-squared value from 0.66 to 0.77 and thus and improvement in accuracy. Often times we would take a log of the variable we are trying to estimate when there is a high level of skewness. This helps normalize the residuals and results in a better model, however, this is not the case for our dataset as we see happiness generally follows a normal distribution as seen in page 3. We utilize four of the six variables in our dataset to interpret future values when determining a happiness score. In terms of simplicity, four out of six variables may not seem like the most straightforward way of constructing this model, however, the World Happiness Report has filtered many variables and indicators in the past decade by narrowing down to just 6 very useful and statistically significant factors. In this context, our model uses a very limited and simplistic structure in a sea of potential variables that could also be used to describe happiness while still remaining accurate when predicting these scores. Thus, our multilinear regression model is

$$Score = 0.8105*gdp + 1.0166*ss + 1.1414*he + 1.8458*f + 1.8921$$

Such that *gdp* is GDP per capita, *ss* is social support, *he* is healthy life expectancy, and *f* is freedom to make life choices.

Prediction

Using this model we are able to predict values of happiness with a 95% prediction interval by using the predict() function built into R. We will use data from 2015 of a distinct country not found in the dataset used in this analysis from 2018. Suriname is a South American country with a GDP per capita of 0.99534, 0.972 for social support, 0.6082 for healthy life expectancy, and 0.5966 for freedom to make life choices. After running the prediction function, we get the following results.

fit	lwr	upr
5.482456	4.389013	6.575898

Thus, we can predict the happiness score of the country Suriname to be within 4.389 and 6.576 with a 95% probability where our predicted happiness score is 5.482. This value is what our model predicts based on 2018 and is slightly different from the actual happiness score value in 2015 of 6.269.

Summary

A happiness score can be predicted for a country using a multilinear regression model which is estimated by GDP per capita, social support, healthy life expectancy, and freedom to make life choices. With an r-squared value of 0.7709, this relationship is shown to be statistically significant after evaluating t-tests for the four variables used. The interpretation for this result follows that 77.09% of the variation in happiness can be explained by this model. National governments can utilize these results to enact new policy and change their country for the well-being of its citizens. It is important to distinguish between correlation and causation for any statistical relationship. The variables with high correlation are not necessarily causing a higher level of happiness within countries. The most we can say at this point is that countries who have a high GDP per capita, social support, healthy life expectancy, and freedom to make life choices are more likely to be happier. With this in mind, we have the ability to do more research, build more models and truly understand what makes a country, and more specifically people, happy. For future studies an intriguing analysis would be to divide countries by the type of government to explore the ideal foundational structure of ideological policies that correlate with a happier population. Some more ideas that were expressed at the beginning of this paper that I think would provide significant results would be to explore city and rural living, religious affiliations, weather, and type of job.