Happiness Project

Sujata Verma

8/11/2020

Table of Contents

INTRODUCTION

In this project, I will be exploring dataset on World Happiness Report 2020 which is based on the global survey of 156 countries. Data is collected through country surveys from 2017-19 conducted by Gallup World Poll. The survey asked repondants to rate their life on multiple parameters on a scale from 0 to 10, with 0 being the worst and 10 being the best. This ranking is called the Cantril Ladder. The six variables are Gross Domestic Product per person, healthy life expectancy, social support, perceived freedom to make choices about one's life, generosity, and perception of corruption. The report defines Dystopia as an imaginary country with the worst possible score in all six variables. The dystopia plus residual factor captures the three-year average lowest score of 1.97 and the unexplained variation from each of the six variables.

In my project, I will add another explanatory variable, income ineqaulity within each country, as measured by the Gini Coefficient. Per-capita GDP captures the economic well being of the residents of a country but the GDP is not equitably distributed within the country. My hypothesis is that when people are asked about their economic well-being, the disparities in income influences their perception about their relative well-being and hence their happiness score.

In order to find out the affect of income inequality in a country on its happiness score, I will combine the World Happiness Report data set with Gini coefficients data from the World Bank for various countries, for the latest year that is available. I will then check for missing values and try and update the data set. Gini coefficient information from the CIA World Factbook will be used to supplement the data for countries where World Bank data is not available. The Gini coefficents will be scaled from 0 (perfect equality) to 1 (perfect inequality).

GOAL OF THE PROJECT AND MAIN STEPS

The central **objective** of this project is to explore what determines happiness of people. More specifically, does income inequality within nations affect the Happiness Scores? I will identify the best machine learning model for predicting the determinants of happiness by fitting a few regression models to the data and calculating the Root Mean Square Error. All models with RMSE of less than 0.80 will be acceptable and the model with the lowest RMSE will be deemed best.

Main steps include creating the combined dataset, data visualization, and fitting models with the best predictive power.

DATA WRANGLING

In this section, the data is imported and observations with missing values are removed.

The first step is to load the libraries and options that will be needed.

The second step is the load the two data-sets, Happiness data from the World Happiness Report 2020 and Income inequality data from the World Bank website and then to combine them into a single data set. The two data files are available on my github page.

Datasets:

https://raw.githubusercontent.com/econverma/The-Happiness-Project/master/WHR20.csv

https://raw.githubusercontent.com/econverma/The-Happiness-Project/master/gini%20updated.csv

The Happiness Report dataset:

country	region	score	log_pcgdp	soc_support	life_exp	freedom	generosity	corruption	dys_res
Finland	Western Europe	7.81	10.6	0.954	71.9	0.949	-0.059	0.195	2.76
Denmark	Western Europe	7.65	10.8	0.956	72.4	0.951	0.066	0.168	2.43
Switzerland	Western Europe	7.56	11.0	0.943	74.1	0.921	0.106	0.304	2.35
Iceland	Western Europe	7.50	10.8	0.975	73.0	0.949	0.247	0.712	2.46
Norway	Western Europe	7.49	11.1	0.952	73.2	0.956	0.135	0.263	2.17
Netherlands	Western Europe	7.45	10.8	0.939	72.3	0.909	0.208	0.365	2.35

Next, the income inequality data from World Bank is imported and and the two datasets are combined, adding gini coefficient for countries as an explanatory variable.

The World Bank data file was updated for missing values with additional information from CIA World Factbook data, obtained from the Wikipaedia page. Also, names of some of the countries were updated so that they were written exactly the same way as the World Happiness report dataset. The Gini coefficents were scaled from 0 (perfect equality) to 1 (perfect inequality). The missing data was removed after performing summary statistics on the combined data set.

country	region	score	log_pcgdp	soc_support	life_exp	freedom	generosity	corruption	dys_res	gini
Finland	Western Europe	7.81	10.6	0.954	71.9	0.949	-0.059	0.195	2.76	0.274
Denmark	Western Europe	7.65	10.8	0.956	72.4	0.951	0.066	0.168	2.43	0.287
Switzerland	Western Europe	7.56	11.0	0.943	74.1	0.921	0.106	0.304	2.35	0.327
Iceland	Western Europe	7.50	10.8	0.975	73.0	0.949	0.247	0.712	2.46	0.268
Norway	Western Europe	7.49	11.1	0.952	73.2	0.956	0.135	0.263	2.17	0.270
Netherlands	Western Europe	7.45	10.8	0.939	72.3	0.909	0.208	0.365	2.35	0.285

```
##    country                                          region       score      
##  Length:155         Sub-Saharan Africa                :39   Min.   :2.57  
##  Class :character   Latin America and Caribbean       :21   1st Qu.:4.73  
##  Mode  :character   Western Europe                    :21   Median :5.51  
##                     Middle East and North Africa      :18   Mean   :5.47  
##                     Central and Eastern Europe        :17   3rd Qu.:6.23  
##                     Commonwealth of Independent States:12   Max.   :7.81  
##                     (Other)                           :27                 
##    log_pcgdp      soc_support       life_exp       freedom     
##  Min.   : 6.49   Min.   :0.319   Min.   :45.2   Min.   :0.397  
##  1st Qu.: 8.32   1st Qu.:0.737   1st Qu.:59.0   1st Qu.:0.713  
##  Median : 9.46   Median :0.826   Median :66.1   Median :0.800  
##  Mean   : 9.29   Mean   :0.808   Mean   :64.4   Mean   :0.782  
##  3rd Qu.:10.26   3rd Qu.:0.905   3rd Qu.:69.1   3rd Qu.:0.879  
##  Max.   :11.45   Max.   :0.975   Max.   :76.8   Max.   :0.975  
##                                                                
##    generosity       corruption       dys_res          gini       
##  Min.   :-0.301   Min.   :0.110   Min.   :0.26   Min.   :0.24  
##  1st Qu.:-0.126   1st Qu.:0.684   1st Qu.:1.64   1st Qu.:0.33  
##  Median :-0.034   Median :0.783   Median :2.03   Median :0.36  
##  Mean   :-0.015   Mean   :0.734   Mean   :1.97   Mean   :0.38  
##  3rd Qu.: 0.085   3rd Qu.:0.849   3rd Qu.:2.35   3rd Qu.:0.43  
##  Max.   : 0.561   Max.   :0.936   Max.   :3.44   Max.   :0.63  
##                                                  NA's   :17    

## [1] 138  11
```

This completes the data wrangling step of importing and combining datasets and removing the missing values. We now have data on 138 countries and 11 variables with no missing values. The dependent variable is the happiness score. Here are the all the variables contained in the final data set.

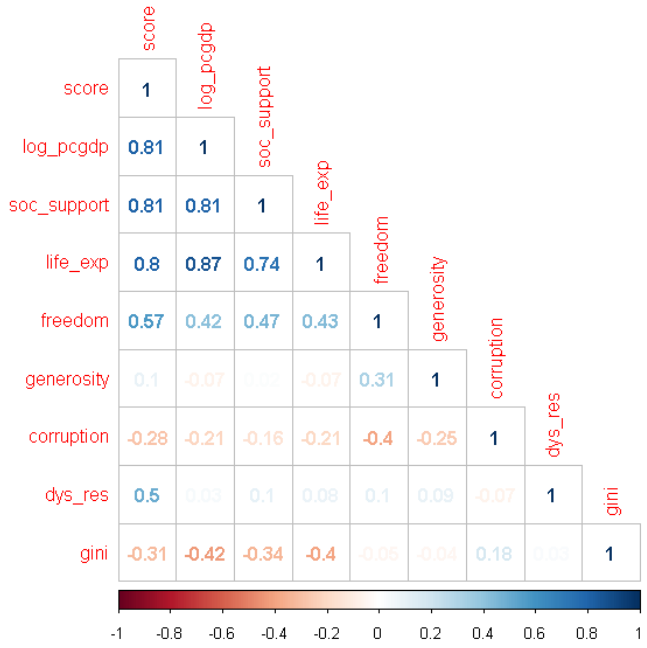
Variable	Explanation
Country	Countries where the surveys were conducted
Region	Region of the world
Score	Overall happiness score from 0 to 10
Log_pcgdp	Per-capita GDP in constant 2011 dollars in purchasing parity parity terms
Soc_support	National average of survey response on presence or absence of social support
Life_exp	Life expectancy at birth
Freedom	National average of survey response on presence or absence of individual freedom
Generosity	National average of survey response to donating to charity
Corruption	National average of survey response to widespread corruption in government or business

Dys_res	Lowest benchmark value plus residual
Gini Coefficient	Income inequality within a country

DATA VISUALIZATION

The following section contains an exploratory data analysis.

Correlation among Variables: Let us take a look at the correlation among the variables by building a correlation plot. It will tell us which are the important explanatory variables affecting the happiness score and whether the explanatory variables have a positive or negative effect on the happiness score. We exclude country name and region for the correlation plot. Note that correlation doesn't necessarily imply cause and effect.



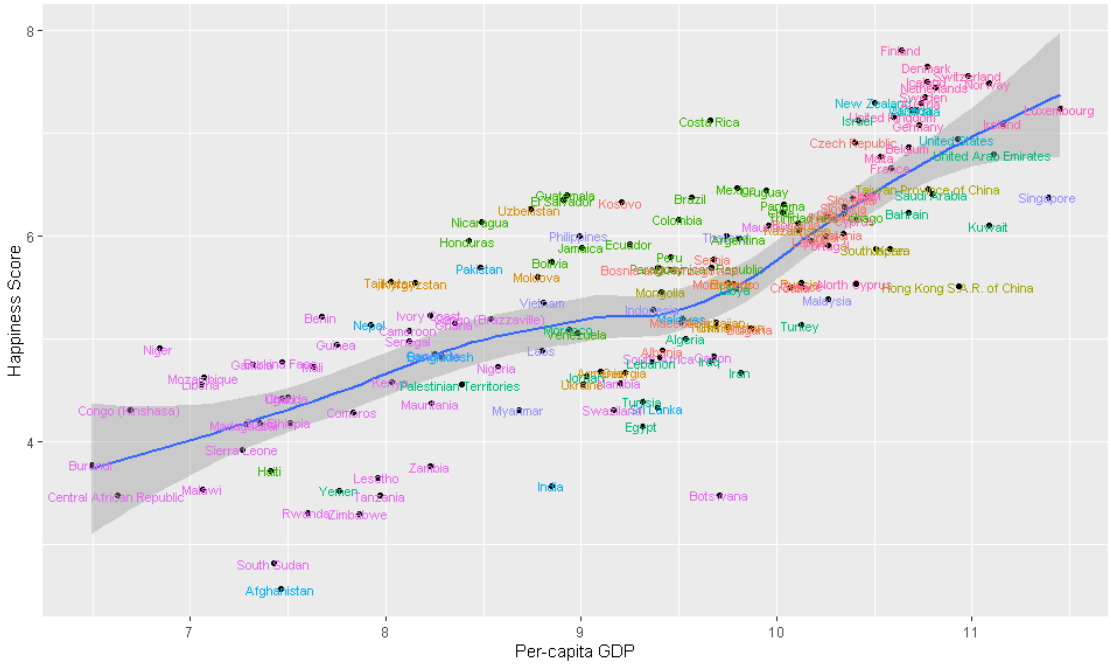
From the correlation plot, we can infer that the Growing Economy, Social Support and Life Expectancy have the biggest positive correlation with the Happiness Score. Freedom to make choices also correlates with Happiness positively. Corruption and Income Inequality have a weak negative correlation with Happiness, while Generosity is not correlated with the Happiness Score. The Dystopia + Residual is positively correlated to Happiness, showing that we are not capturing all of the determinants of happiness.

Other interesting strongly positive correlations are between the Growing Economy and Life Expectancy and also between the Growing Economy and Social Support. Gini coefficient has a weak negative correlation with Per-capita GDP, Social Support as well as Life Expectancy.

Country-wise Scatter Plot

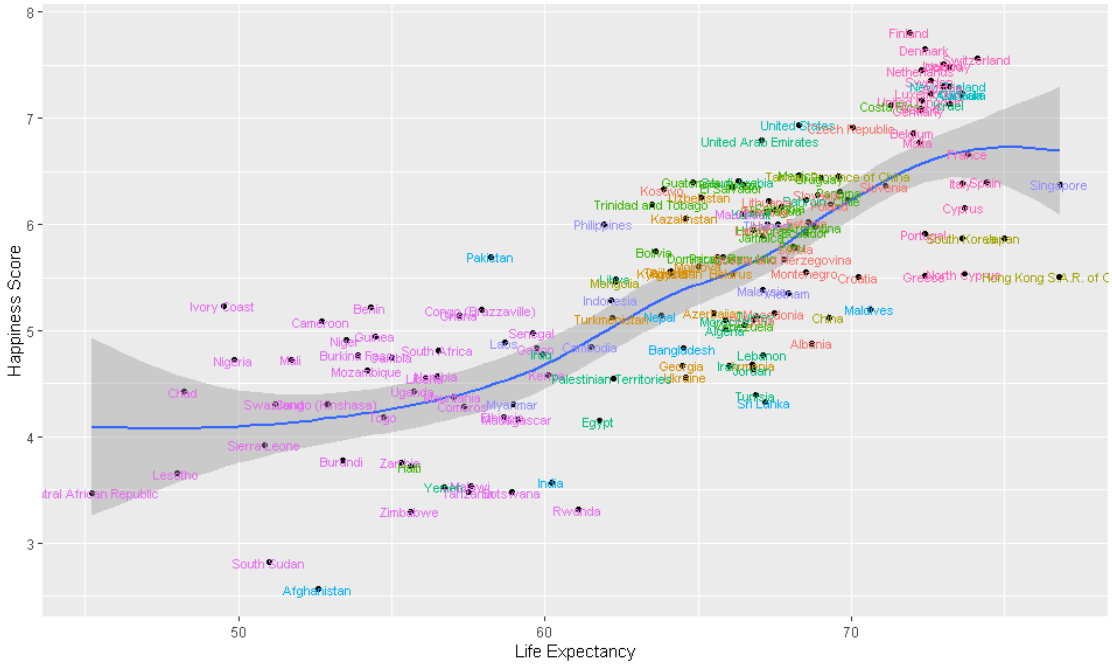
These relationships among Happiness Score and each of the seven explanatory variables is further explored by a country-wise scatter plot for each of the variables against the dependent variable, Happiness Score. Instead of a linear regression line, Loess method is used to get a smooth curve. Each region is distinguished by a separate color. Note,we don't explore Dystopia + Residuals in the future analysis because it is not a factor causing happiness or unhappiness and we focus on what makes people happy.

```
Economy and Happiness
```



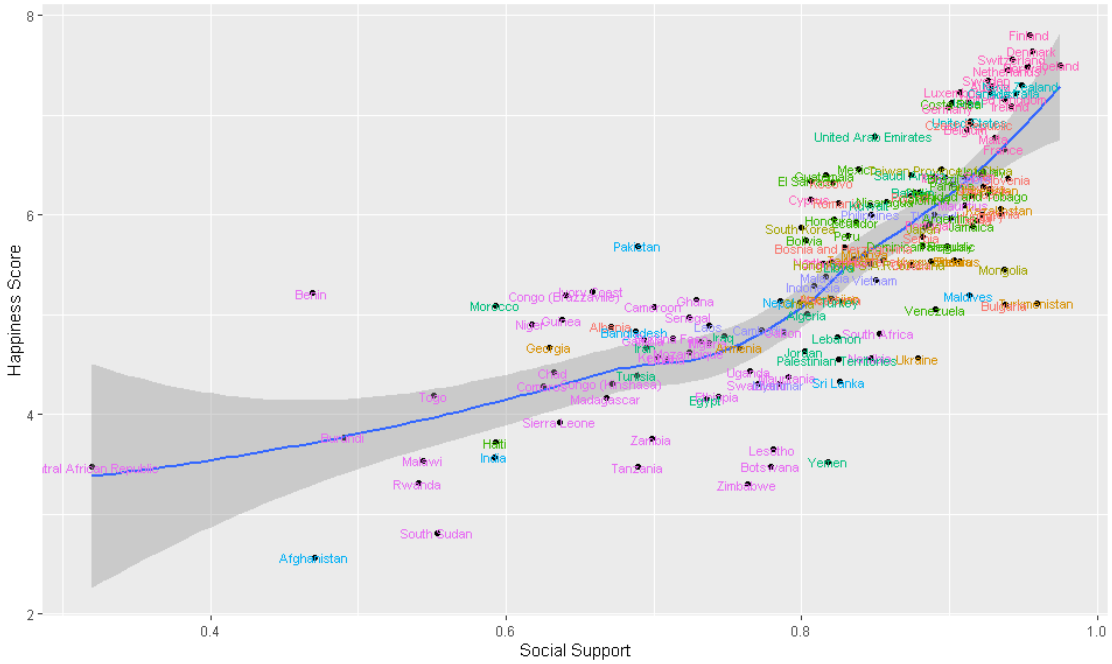
The above graph confirms that wealthier the country, more is the reported happiness.

```
Life Expectancy and Happiness
```



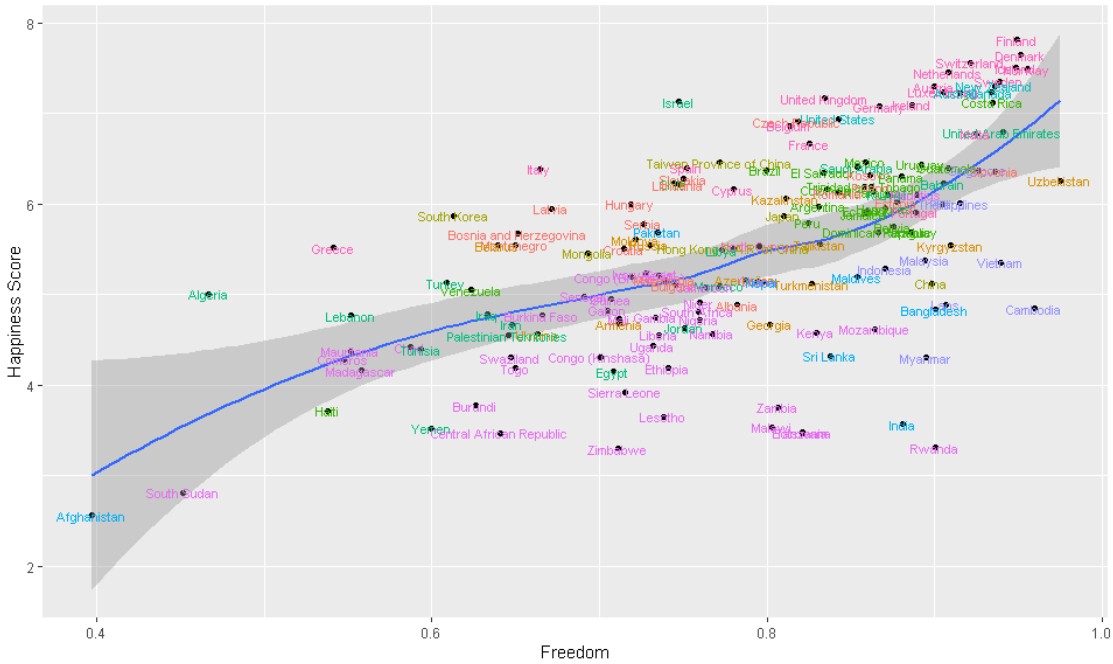
The healthier the population of a country as denoted by life-expectancy, more is the reported happiness. The curve levels off at very low and very high life-expectancy rates.

```
Social Support and Happiness
```



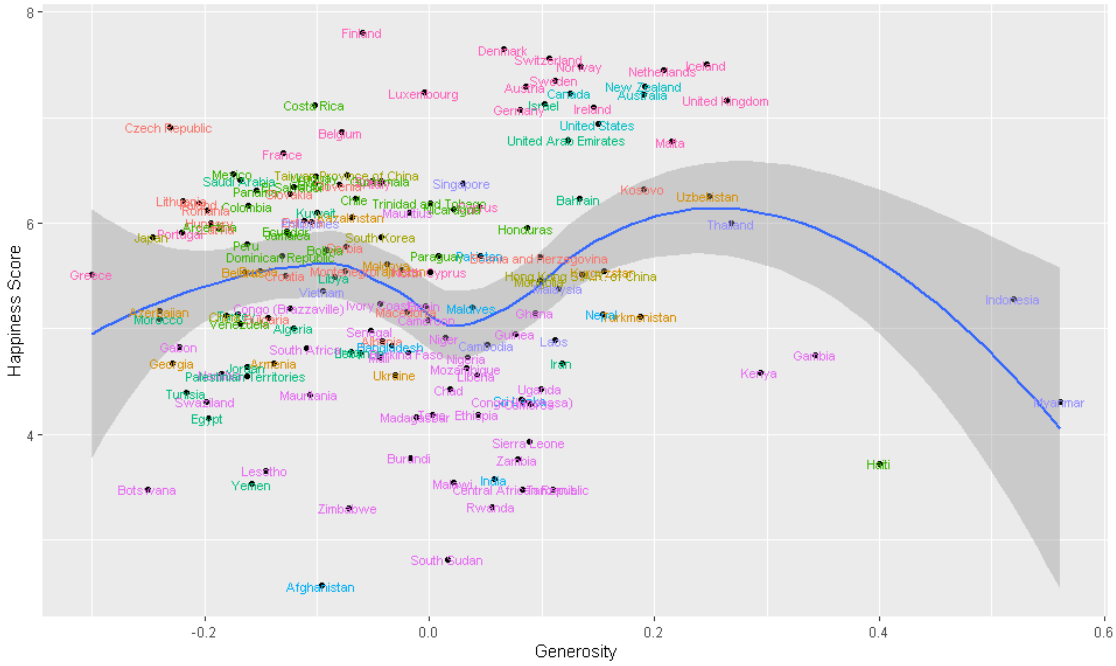
The relationship between friends and family support and happiness is a very strong one. The happiness increases exponentially as the family support increases. Western European countries and Australia and New Zealand ranked at the top in terms of social support. Interestingly, Turkmenistan and Slovenia ranked in the top 10 in terms of family and social support.

```
        Freedom and Happiness
```



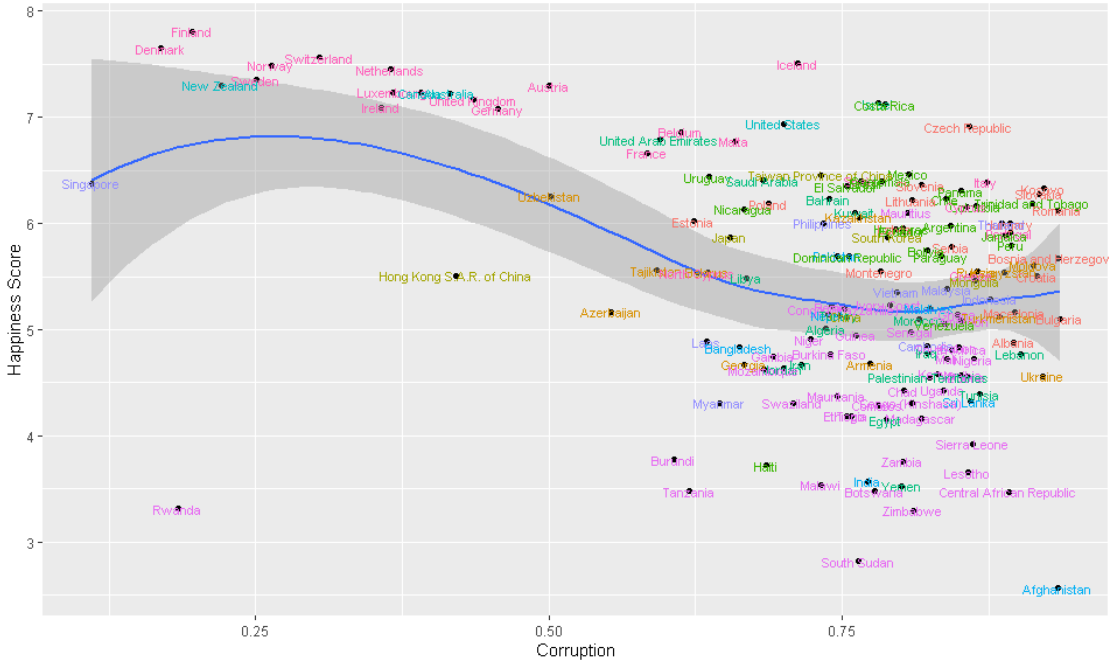
The chart above shows that happiness is higher in countries where residents believe that they have the freedom to make decisions about their life.

```
Generosity and Happiness
```



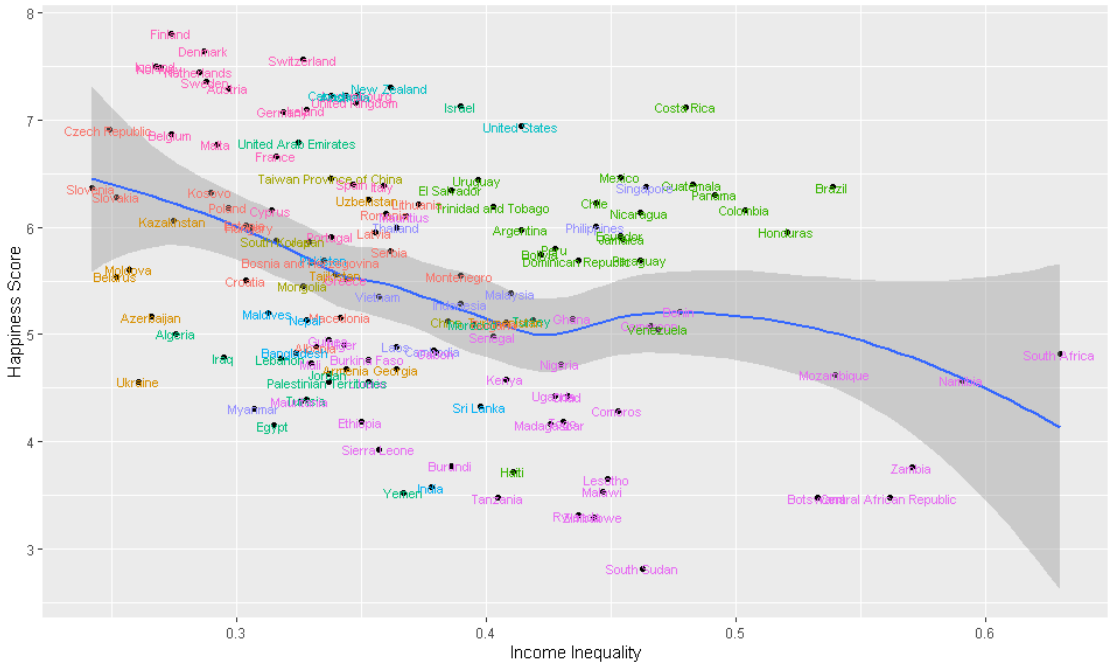
Being charitable and generous doesn't make us happier, according to the above chart. The three most generous countries were Myanmar, Indonesia and Haiti, yet the people in these countries are not among the richest or the happiest.

```
Corruption and happiness
```



There is a negative relationship between perceptions of corruption and happiness. What is interesting about this chart is the preponderance of multiple counties at a very high level of corruption. It would seem that corruption is a way of life in most of the world.

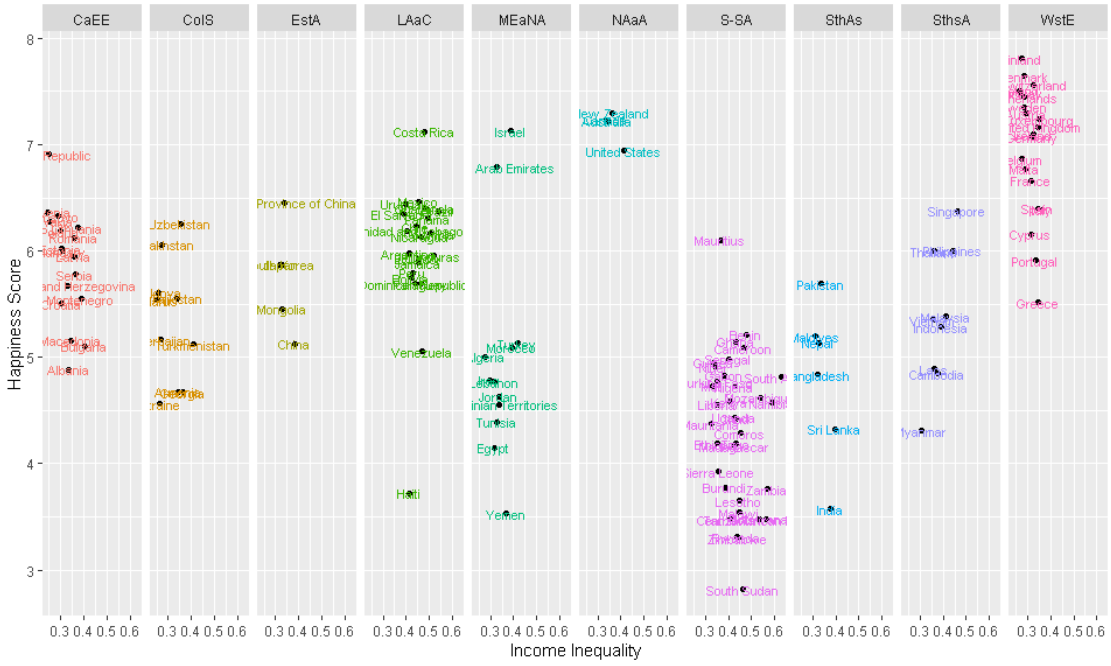
```
        Income Inequality and Happiness
```



There is a negative relationship between Income Inequality and Happiness. This chart supports the hypothesis that income disparities lowers happiness.

```
      Regional Income Inequalities and Happiness:
```

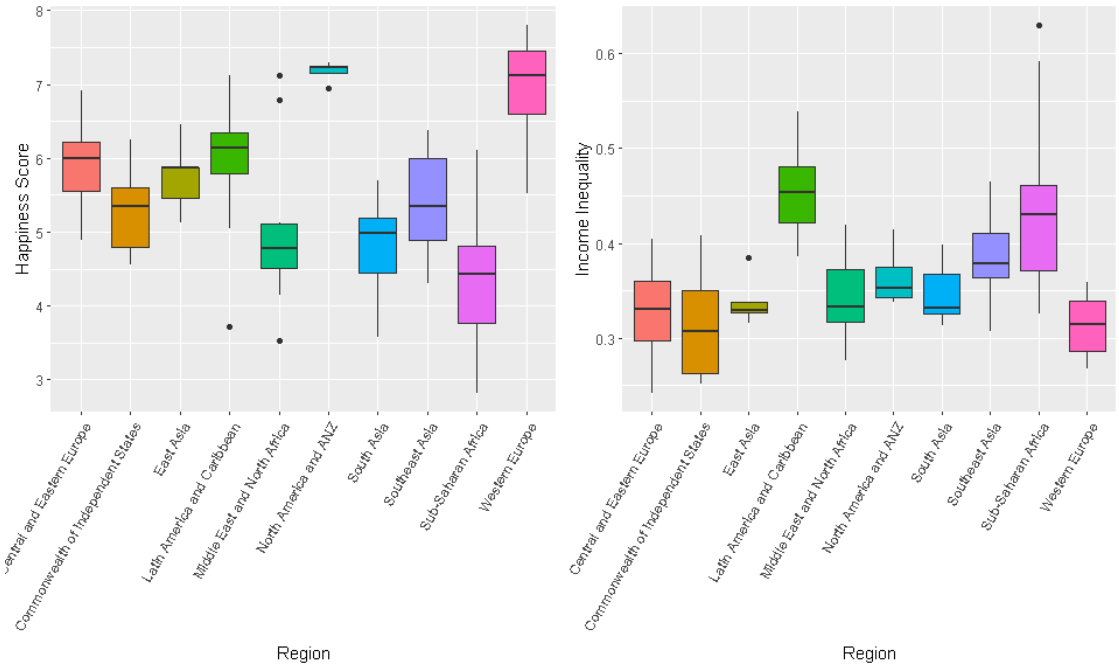
In this section, we will explore the link between happiness and income inequality by region. Let us first take a look at the scatter plot of happiness Score, but this time dis-aggregating by region. We abbraviate the ten regions as follows: Central and Eastern Europe(CaEE), Commonwealth of Independent States(CoIS) , East Asia(EstA), Latin America and Caribbean(LAaC), Middle East and North Africa(MEaNA), North America and ANZ(NAaA), Sub-Saharan Africa (S-SA), South Asia(SthAs), Southeast Asia (SthsA), and Western Europe (WstE).



We can see from the above chart, that Sub-saharan Africa and Latin America and the Carribean have a higher Inequality and a lower Happiness Score. While North America, Australia and New Zealand as well as Western Europe have a lower Income Inequality and higher Happiness Score.

Instead of scatter plots, we can have a clearer picture of the relationship between region-wise Income Ineqalities and Happiness by creating box-plots.

```
Box-plots of Regional Income Inequality and Happiness
```



From the figure on the left, we can infer that Western Europe and North America, Austraia and NewZealand are the happiest regions. Sub-Saharan Africa, Middle East and North Africa and South Asia score less on the Happiness Score.

From the figure on the right, Latin America and Caribbean and Sub-Saharan Africa are some of the most inequitable regions and Western Europe is the one of the most equitable regions.

Latin America and Caribbean is an exception to this inverse relationship as it has high Happiness Score despite having a high degree of Income Inequality.

METHOD AND ANALYSIS

In the following section, various machine learning models are built to predict the happiness score. Each model is built by using the caret package, and applying it to our combined data set.

For each model, the following steps are taken:

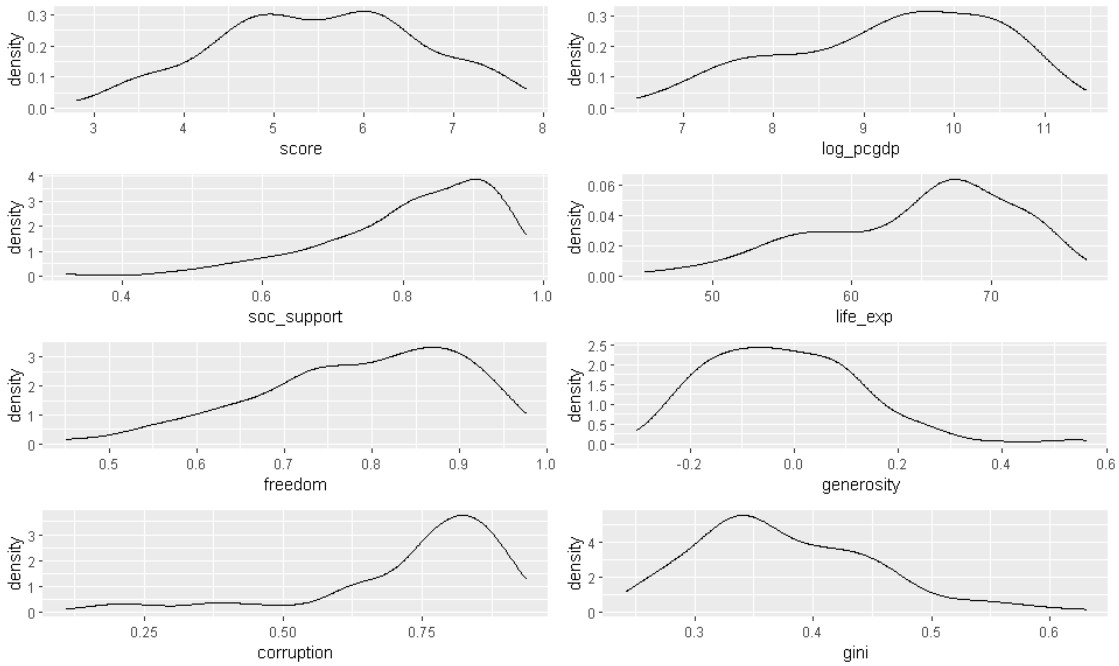
1. Training the model using the training set
2. Making predictions using the test set
3. Determining the Root mean Square Error of the predictions
4. Storing the results of the model in a results table

The best model will be the one with the lowest Root Mean Square Error (RMSE), which calculates how far is the actual happiness score from the score predicted by the model.

First step is to create training and testing data sets. I decided to split the entire data set containing 138 rows equally between the training and testing data-sets, since the number of observations is not very large, so an uneven split will make the results for the smaller data set very skewed.

Which Regression Model will be appropriate?

We have one dependent variable, happiness score, and seven explanatory variables. But do these variables follow a normal distribution? To find out we create smooth density functions of each of these variables below.



As we can see from the above chart, the variables don't show a normal distribution. Instead of a multivariate normal model, it will be more appripriate to apply a generalized linear regression model (GLM).

Generalized Linear Regression Model

```
##
## Call:
## glm(formula = score ~ ., data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4801  -0.2790   0.0651   0.3671   0.8510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.8218     1.0858   -1.68    0.099 .
## log_pcgdp     0.2712     0.1283    2.11    0.039 *
## soc_support   1.2384     1.1579    1.07    0.289
## life_exp      0.0491     0.0198    2.48    0.016 *
## freedom       1.9576     0.8644    2.26    0.027 *
## generosity    0.4876     0.5518    0.88    0.380
## corruption   -1.1696     0.5377   -2.18    0.034 *
## gini          0.0104     1.0847    0.01    0.992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.33)
##
##     Null deviance: 80.043  on 67  degrees of freedom
```

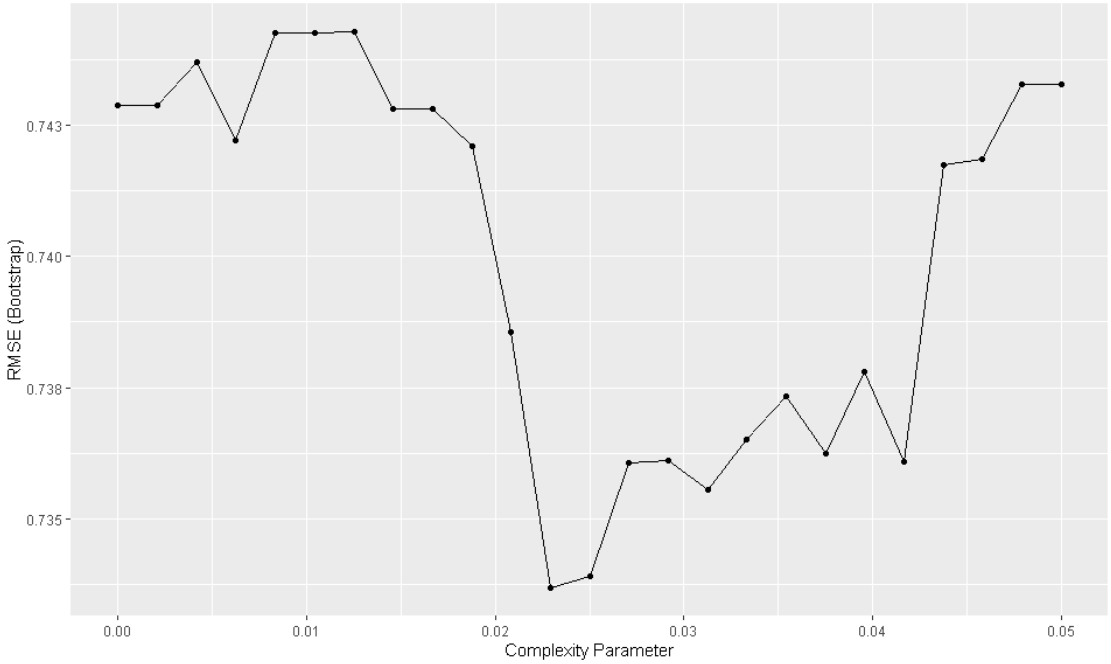
```
## Residual deviance: 19.795  on 60  degrees of freedom
## AIC: 127.1
##
## Number of Fisher Scoring iterations: 2
```

Model	RMSE
GLM Model	0.611

The GLM Model shows some interesting results. The RMSE is low enough to accept the model. The most significant explanatory variables are Per capita GDP, Life expectancy,Freedom and Perceptions of Corruption.

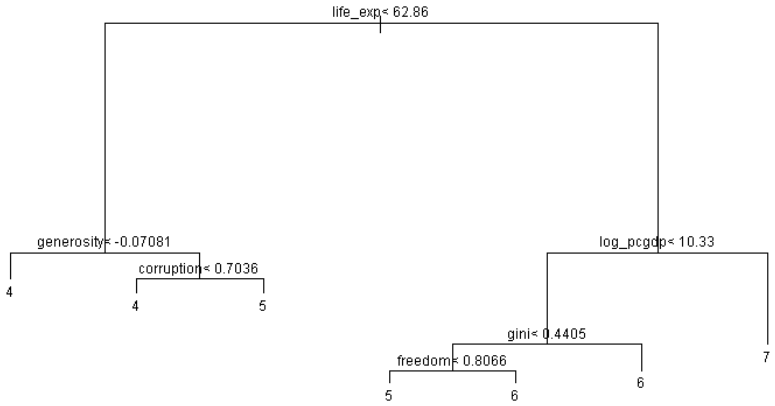
Decision Tree Model

As a second model, we will try and fit the continuous variable decision tree model(DT)




```
##        cp
## 12 0.0229
```

Model	RMSE
GLM Model	0.611
Decision Tree Model	0.731

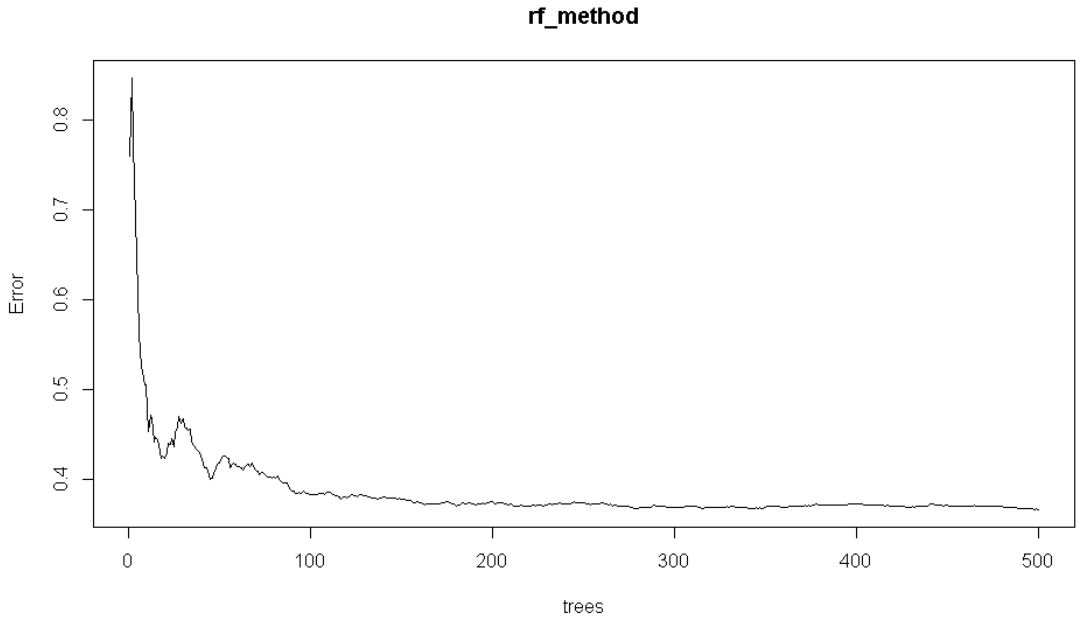


The RMSE is low, though not as low as the GLM Model, is still low enough for us to accept the model. The decision tree tells us that happiness score of 7 is possible when Life Expectancy and Per-capita GDP are high. Similarly, lower Life Expectancy and lower Generosity scores can lead to a low Happiness Score of 4.

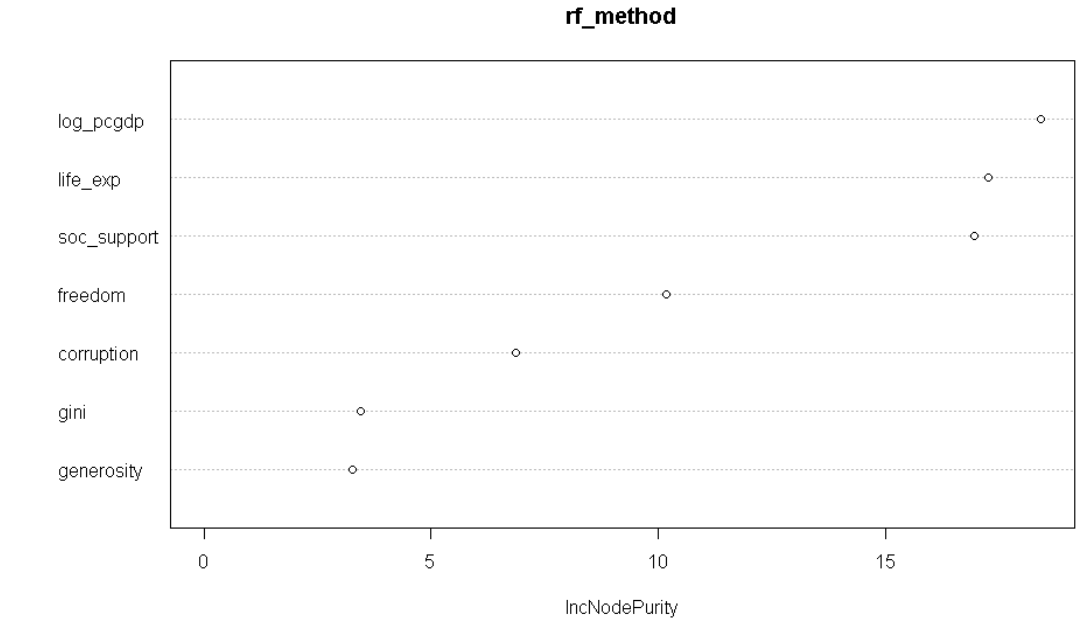
An advantage of this model is the ease of interpretation and a couple of different scenarios about what variable mix can lead to a higher or lower happiness score.

Random Forest Model

Lastly we fit the Random Forest Modelto out training and testing datasets.




```
##              IncNodePurity
## log_pcgdp            18.42
## soc_support          16.95
## life_exp             17.27
## freedom              10.17
## generosity            3.28
## corruption            6.86
## gini                  3.44
```



Model	RMSE
GLM Model	0.611
Decision Tree Model	0.731
Random Forest Model	0.507

We can infer from the figure above, that at about 200 trees, the error cannot be reduced any further.

The model also indicates that Per-capita GDP, Social Support and Life Expectancy are the most important determinants of happiness.

The RMSE obtained from fitting this model is low enough for the model to be acceptable. The Random Forest Model has a lower RMSE than the GLM Model.

RESULTS AND CONCLUSIONS

Model	RMSE
GLM Model	0.611
Decision Tree Model	0.731
Random Forest Model	0.507

The central **objective** of this project was to explore the determinants of happiness and to determine the best machine learning model to predict happiness.

In terms of the fitting a predictive model, all three models considered, the GLM Model, the Decision tree Model and the Random Forest Model, fitted very well with the RMSE below the target of 0.8.

The Random Forest Model did the best as it gave us the lowest RMSE.The variable importance obtained from this model pointed to Per capita GDP,Life Expectancy and Social Support as the three most important determinants of happiness. Income inequality was not among the most important determinants.

LIMITATIONS AND FURTHER EXPLORATION

The data set was quite small since there are only 138 rows, one for each country. Splitting the data into training and testing sets further reduced the accuracy of the models.

More explanatory variables, for example, Crime Statistics and Quality of Environment in a country, can be explored to explain happiness of people.

Also, more analysis can be done regarding inter-dependence among explanatory variables, as we found in the Correlation Plot.

REFERENCES

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network

The World Bank https://data.worldbank.org/indicator/SI.POV.GINI

CIA World Factbook

https://en.wikipedia.org/wiki/List_of_countries_by_income_equality