World Happiness Report

Abdallah Hazem

5/16/2020

Introduction

Predictive machine learning algorithms are often at the forefront of modern computational science. Autonomously estimating results given troves of simulated data is one of the most powerful uses of machine learning. Examples vary from determining exoplanet existence based on light flux data to leveraging the flow of the US bond and stock markets. Depending on the type of data you are processing, prediction can be accomplished through classification models, random forest, k-nearest neighbors or many other machine learning algorithms. This project employs regression analysis in order to study the happiness of countries.

Overview

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

The World Happiness Report, uploaded to Kaggle by the Sustainable Development Solutions Network, is described as "a landmark survey of the state of global happiness." The most recent survey was conducted throughout 2019 and ranks all countries with a happiness score. Particular factors of a society are included with the score that help gauge the livelihood in that country.

The rest of the report is organized as follows: second, the data set is dissected and visualized in the **Data** section. Third, the data is partitioned and the three modeling methods are shown in the **Methods** section. Fourth, the three models are compared in the **Results** section. Finally, the **Conclusion** section summarizes and compares the results.

Data Exploration

The data being used has 156 countries listed with 9 informative columns. These columns are happiness rank, country name, overall happiness score, and six factors that aid in defining the happiness score: GDP per capita, Social Support, Life Expectancy, Freedom, Generosity, and Truth.

The scores are based on answers from the Gallup World Poll to the main life evaluation question asked in the poll. This question, known as the *Cantril ladder*, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The emblematic Orwellian dystopia, a society in which the above aspects are destructive to the

welfare of a free and open society, is most closely associated with the lowest scoring countries. Therefore, higher scores represent, essentially, a closer approximation to a Utopia.

Confusion exists as to how the numbers for each factor apply to the happiness score (discussions can be found (https://www.kaggle.com/unsdsn/world-happiness/discussion/116054) and (https://www.kaggle.com/unsdsn/world-happiness/discussion/35141)). These values represent the effect that each societal variable adds to a standard dystopian score. Similarly, having higher factor scores will lead to a higher happiness score. It is worth mentioning that previous versions of the happiness report have a dystopia residual that represent more general improvements to the standard dystopian. This is not available in the most recent (2018, 2019) data sets.

It is important to note that the factor values themselves are not *directly* correlated with the overall happiness score. The goal of this report is to determine the extent in which they are correlated. The previously mentioned discussions state the best model one can use to estimate happiness scores is the sum of the individual factors, that is `sum(data$[factors]) = data$Score`. This model is presented along with a generalized linear model equation based on 2019 data alone and 2018/2019 data together.
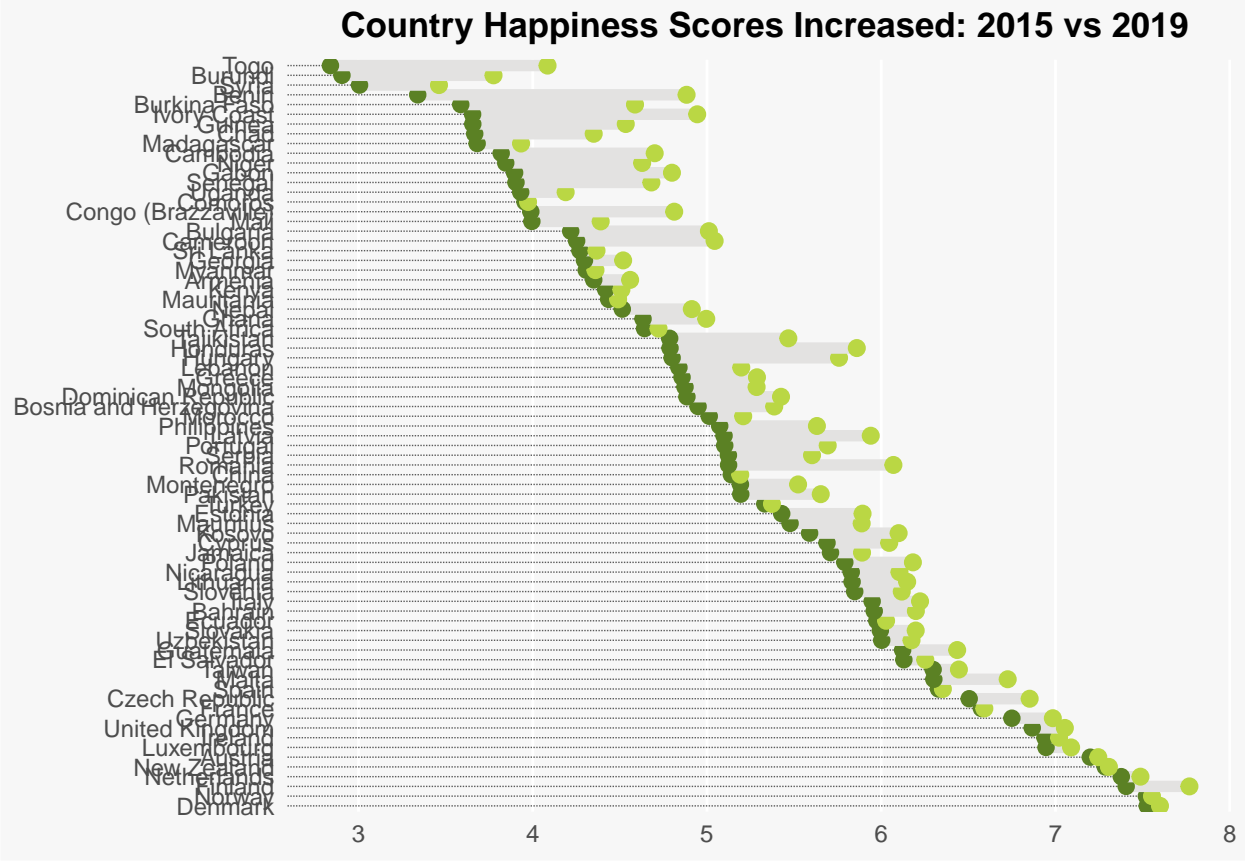
Analysis

Visualization

In this section, we will play with different variables to find out how they correlate with each other.
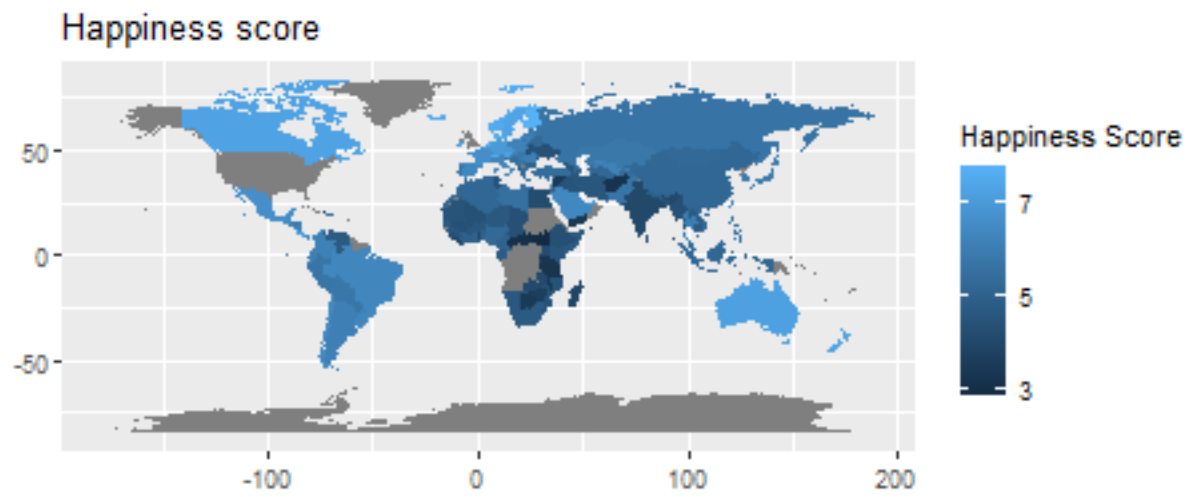
The qualitative structure and the general relationship of the data is described above. The data found here is quantitatively normal. A histogram of the happiness scores shows an approximately normal distribution and descriptive statistics support this claim.
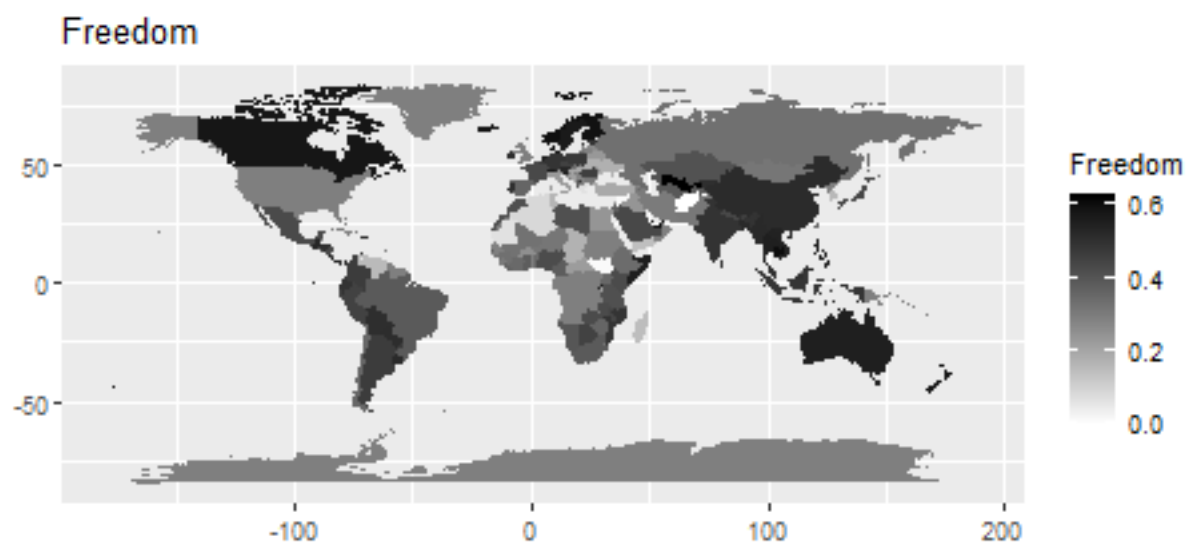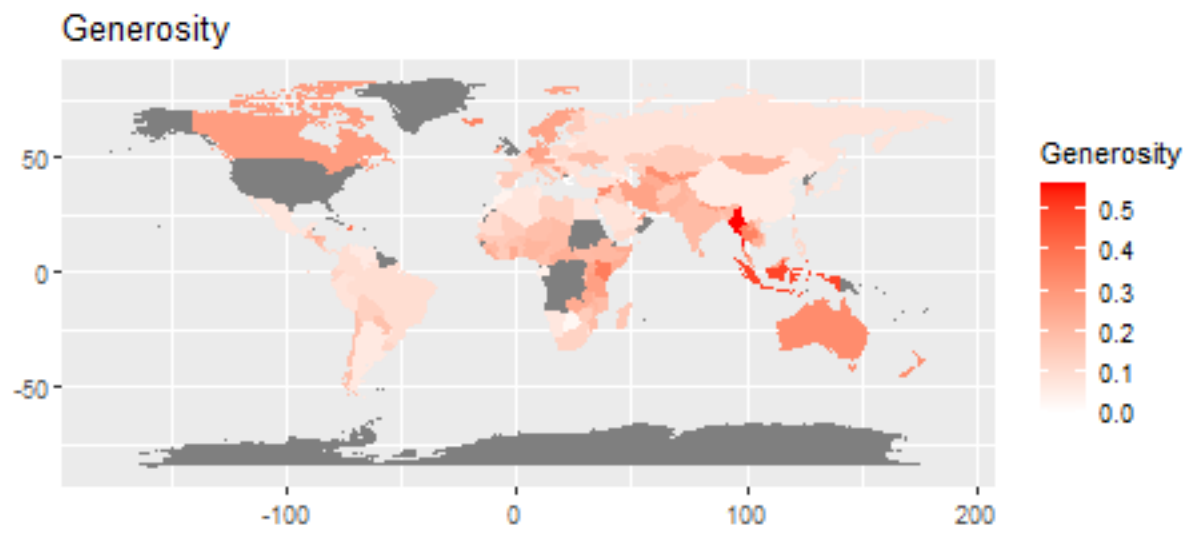

2019 Happiness Scores
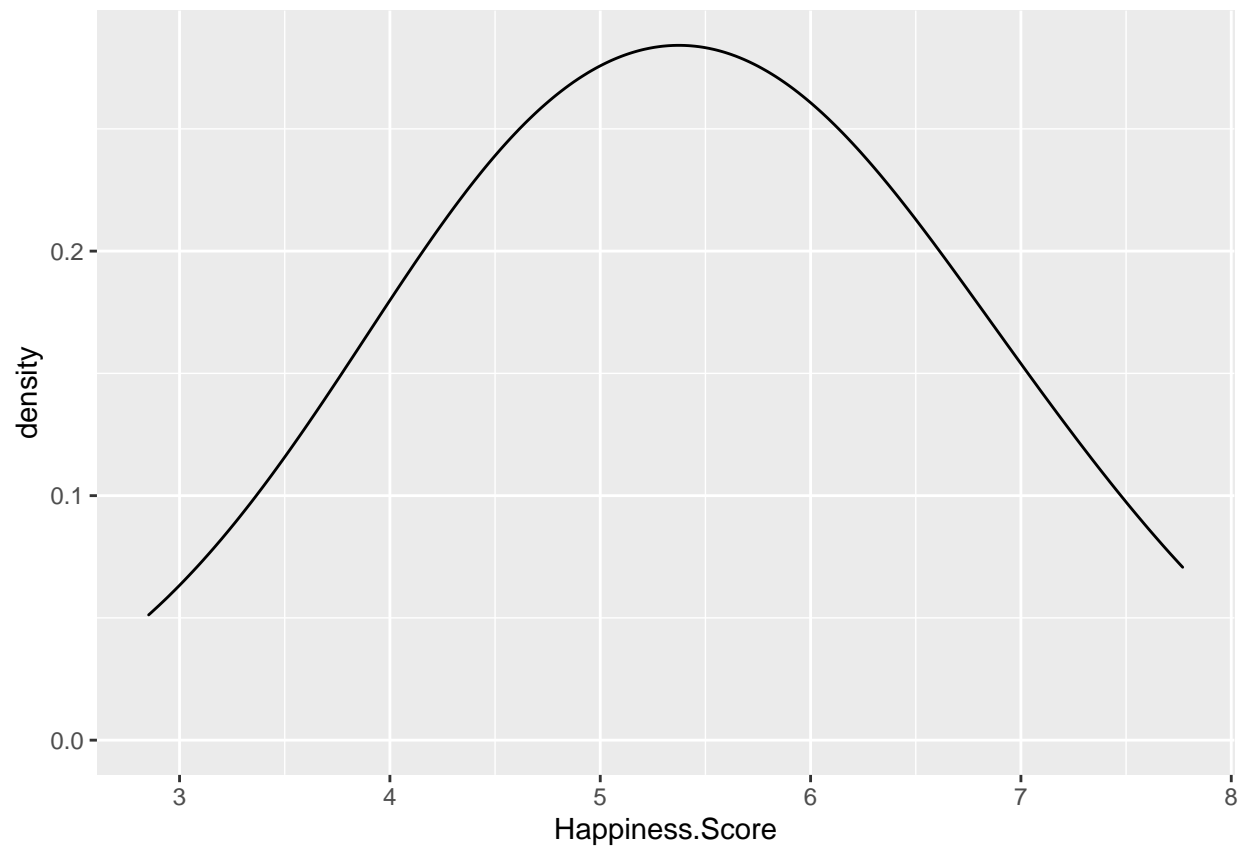
The histogram is right-leaning. This can be shown quantitively by comparing the mean score (5.4070962) and the median score (5.3795). A mean greater than the median signifies that there are more values larger that the middle. The range of scores is (2.853, 7.769). The highest and lowest observations are displayed below:

Country Happiness Scores Increased: 2015 vs 2019

Happiness score

Freedom

Trust

Economy

Life.Expectancy

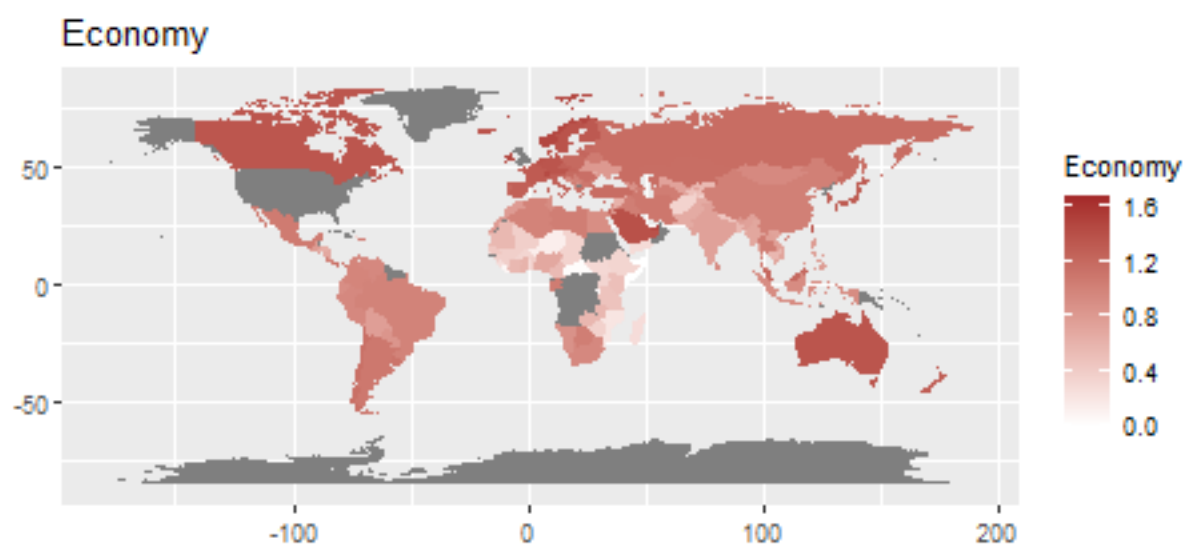## Average value of happiness variables for different continents
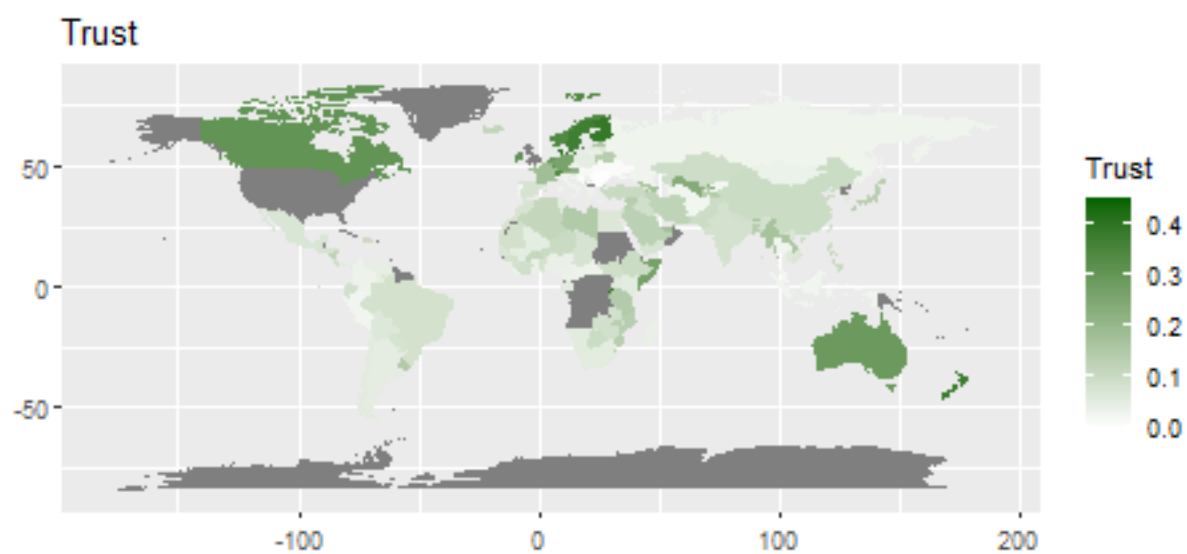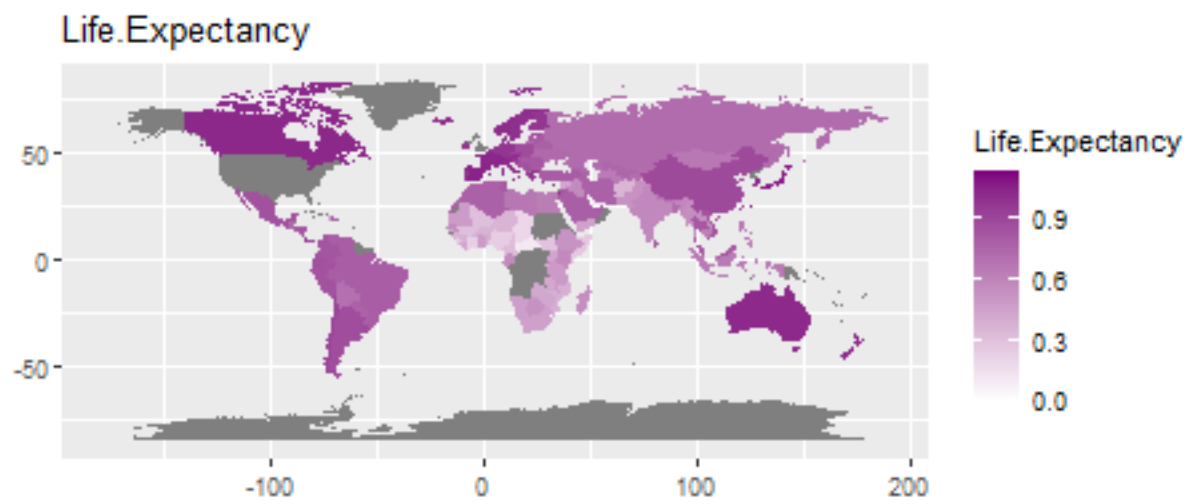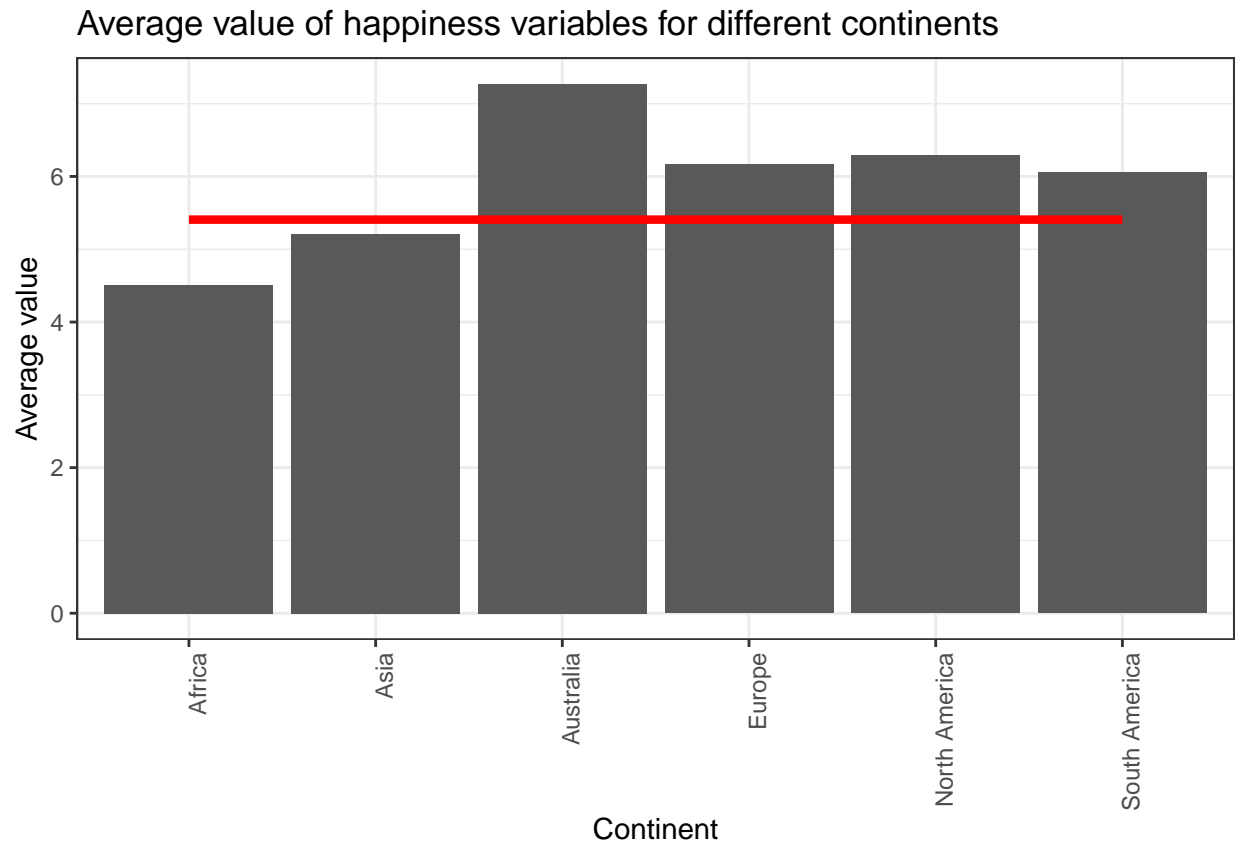
Top 3 happiest regions based on average happiness score are: Australia & North America & Western Europe
It is important to note that both of the first two regions include only 2 countries and Europe has 21 countries.
Additionally all of these regions include countries with developed economies.

The unhappiest region is Africa, which includes 40 different countries. The second unhappiest region is Asia.

Correlation plot

Let's see the correlation between numerical variables in our dataset.
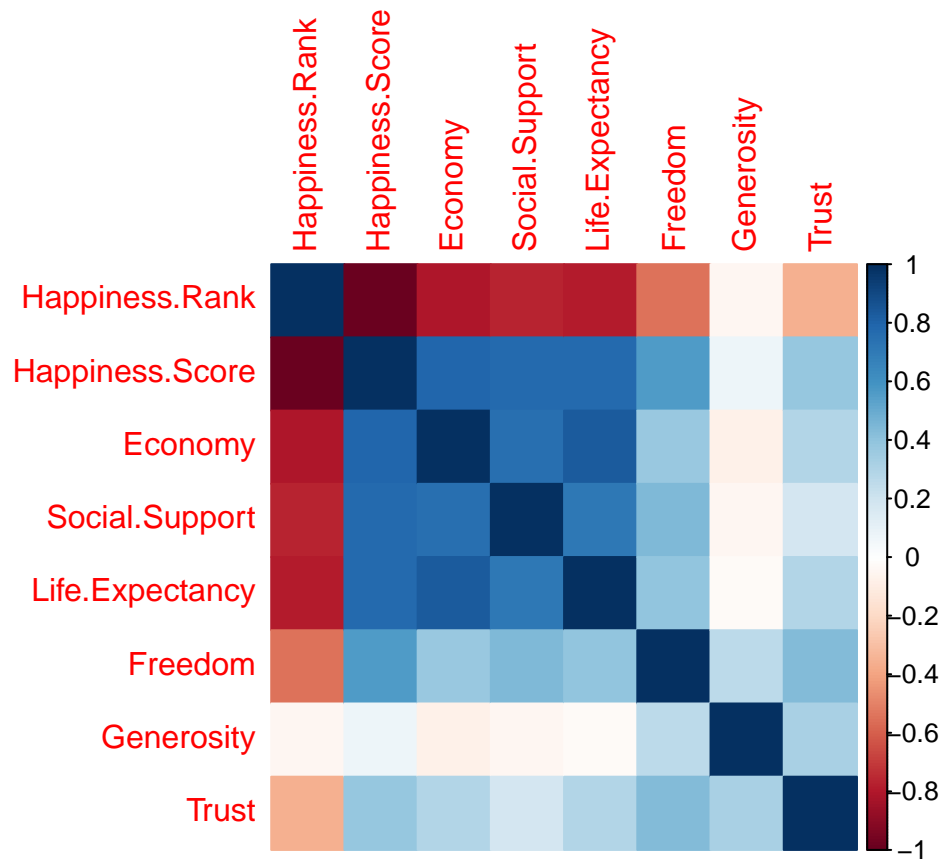
Obviously, there is an inverse correlation between "Happiness Rank" and all the other numerical variables. In other words, the lower the happiness rank, the higher the happiness score, and the higher the other seven factors that contribute to happiness. So let's remove the happiness rank, and see the correlation again.

Comparing different continents regarding their happiness variables

Let's calculate the average happiness score and the average of the other seven variables for each continent. Then melt it to have variables and values in separate columns. Finally, using ggplot to show the difference between continents.

Average value of happiness variables for different continents
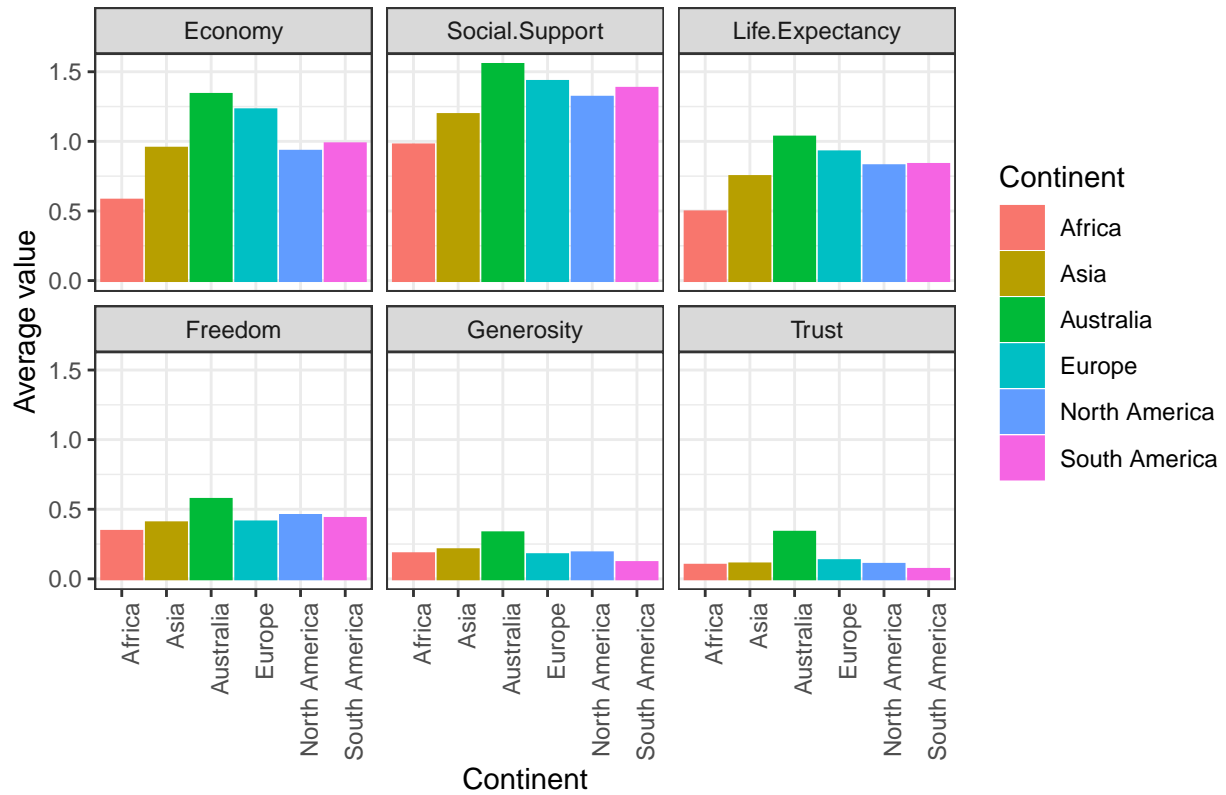
We can see that Australia has approximately the highest average in all fields except dystopia residual, after that Europe, North America, and South America are roughly the same regarding happiness score and the other seven factors. Finally, Asia and Africa have the lowest scores in all fields.

Correlation plot for each continent

Let's see the correlation between variables for each continent.

Happiness Matrix for Africa

	Generosity	0.02	−0.22	−0.35	−0.05	0.25	0.15
		Happiness.Rank	−0.56	−0.71	−0.58	−0.30	0.06
			Social.Support	0.71	0.48	0.23	−0.10
				Economy	0.70	0.20	−0.01
					Life.Expectancy	0.16	0.18
						Freedom	0.35
							Trust
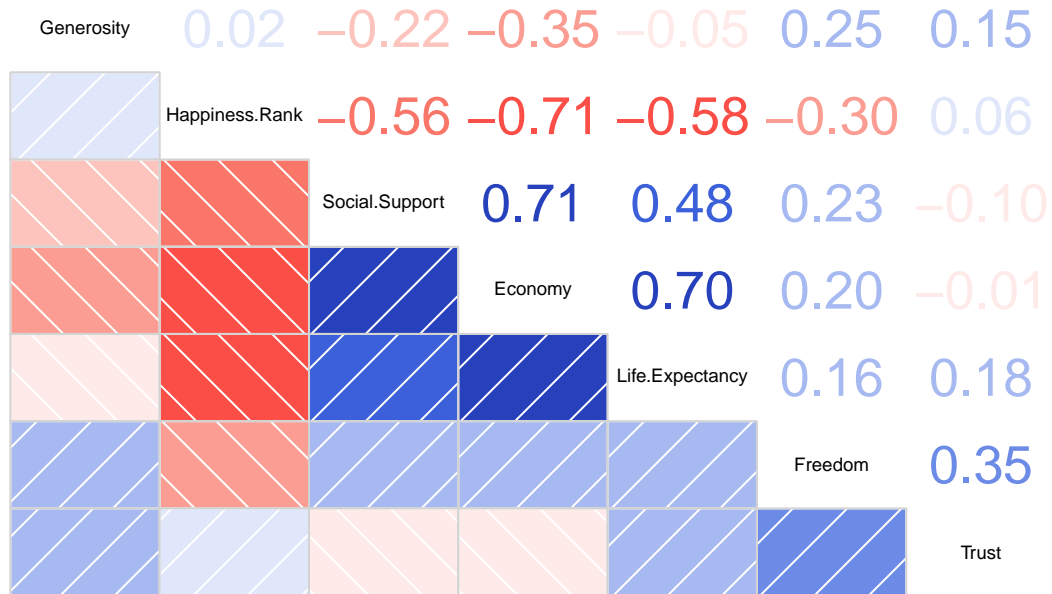
Correlation between "Happiness Score" and the other variables in Africa:
Economy > Life.Expectancy > Freedom
There is no correlation between happiness score and trust.
There is an inverse correlation between happiness score and generosity.

Happiness Matrix for Asia

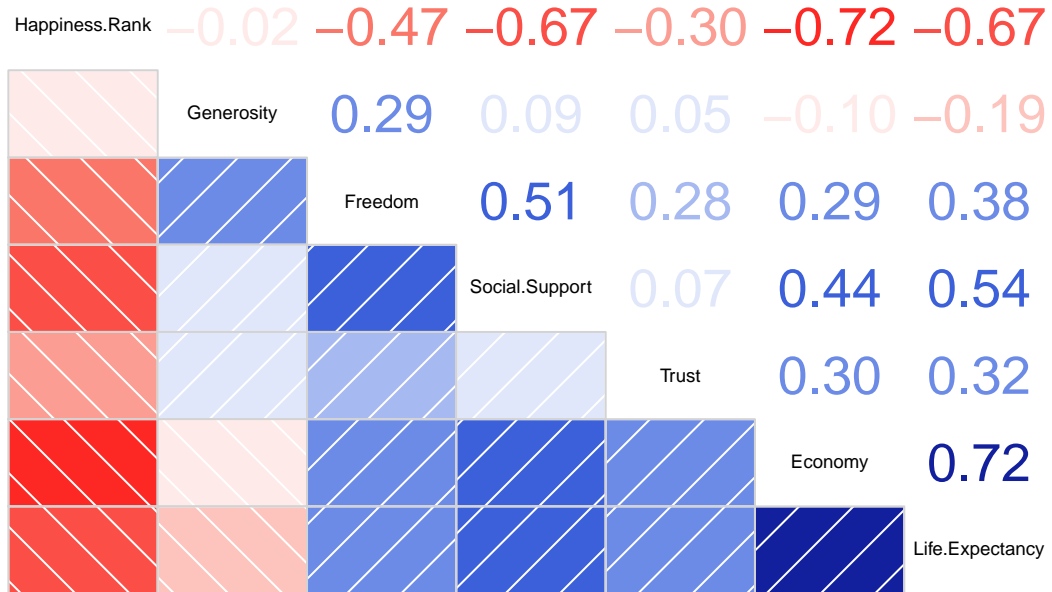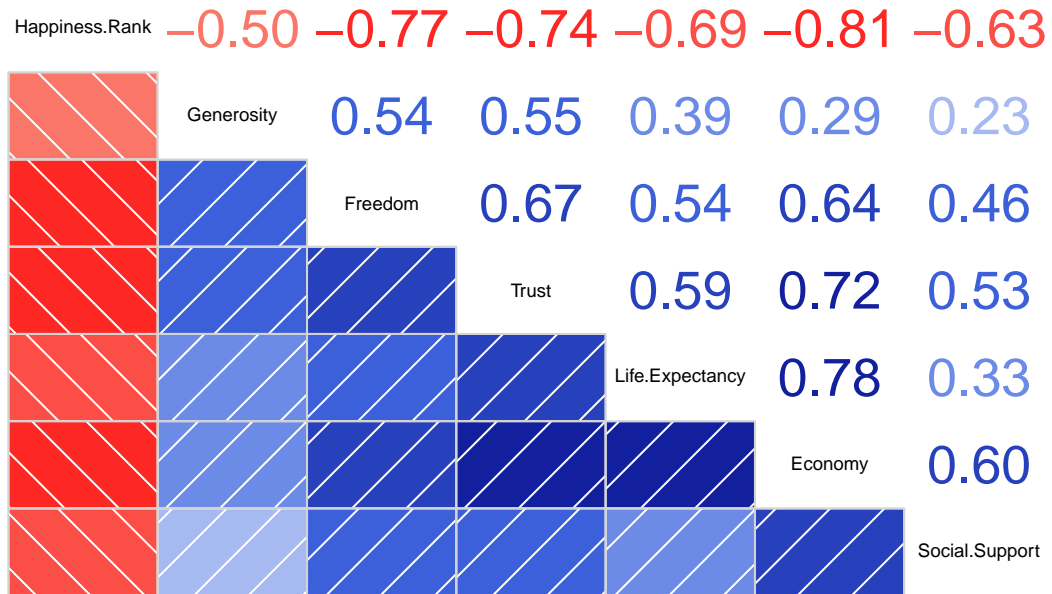|						
Happiness.Rank	−0.02	−0.47	−0.67	−0.30	−0.72	−0.67
	Generosity	0.29	0.09	0.05	−0.10	−0.19
		Freedom	0.51	0.28	0.29	0.38
			Social.Support	0.07	0.44	0.54
				Trust	0.30	0.32
					Economy	0.72
						Life.Expectancy

Correlation between "Happiness Score" and the other variables in Asia:

Economy > Life.Expectancy > Freedom > Trust There is no correlation between happiness score and generosity.

Happiness Matrix for Europe

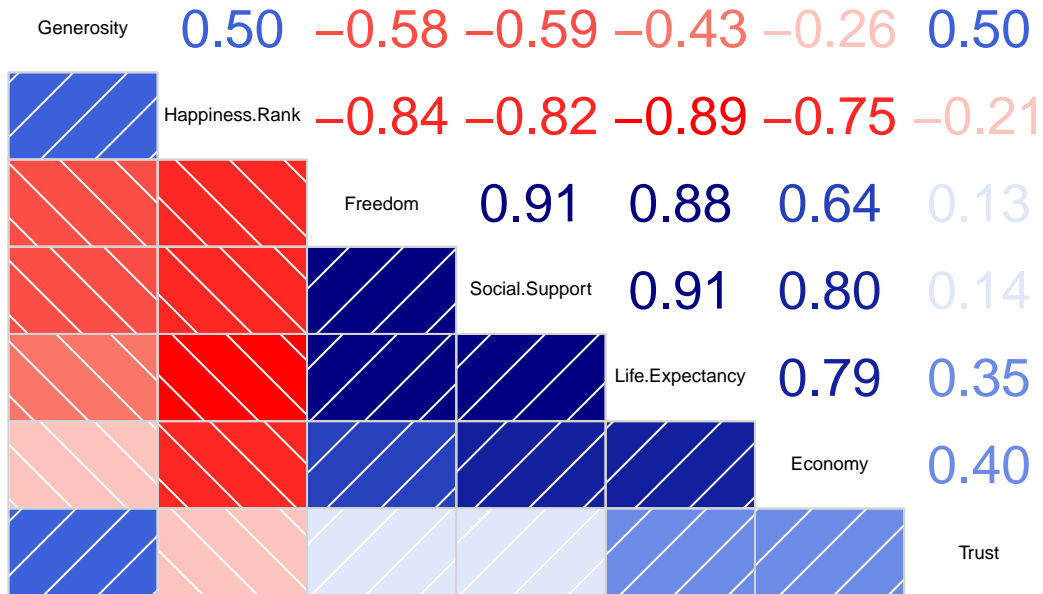|	Happiness.Rank	Generosity	Freedom	Trust	Life.Expectancy	Economy	Social.Support
Happiness.Rank		−0.50	−0.77	−0.74	−0.69	−0.81	−0.63
Generosity			0.54	0.55	0.39	0.29	0.23
Freedom				0.67	0.54	0.64	0.46
Trust					0.59	0.72	0.53
Life.Expectancy						0.78	0.33
Economy							0.60
Social.Support							

Correlation between "Happiness Score" and the other variables in Europe:
Freedom > Trust > Economy > Life.Expectancy > Generosity
The highest correlation between generosity and happiness score took place in Europe.

Happiness Matrix for North America

| Generosity | 0.50 | −0.58 | −0.59 | −0.43 | −0.26 | 0.50 |
Happiness.Rank		−0.84	−0.82	−0.89	−0.75	−0.21
Freedom			0.91	0.88	0.64	0.13
Social.Support				0.91	0.80	0.14
Life.Expectancy					0.79	0.35
Economy						0.40
Trust						

Correlation between "Happiness Score" and the other variables in North America:
Life.Expectancy > Economy > Freedom > Trust
There is an inverse correlation between happiness score and generosity.

Happiness Matrix for South America
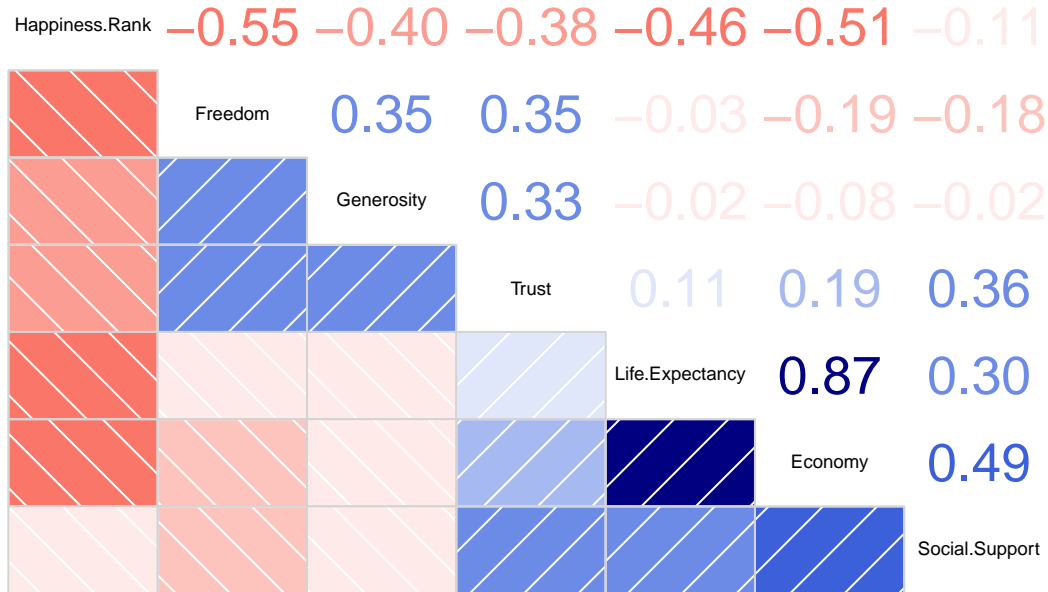
Correlation between "Happiness Score" and the other variables in South America:
Dystopia.Residual > Economy > Life.Expectancy > Freedom > Generosity > Trust > Family
The family is the least significant factor in South America.

## Happiness score comparison on different continents

We will use scatter plot, box plot, and violin plot to see the happiness score distribution in different countries, how this score is populated in these continents and also will calculate the mean and median of happiness score for each of these continents.
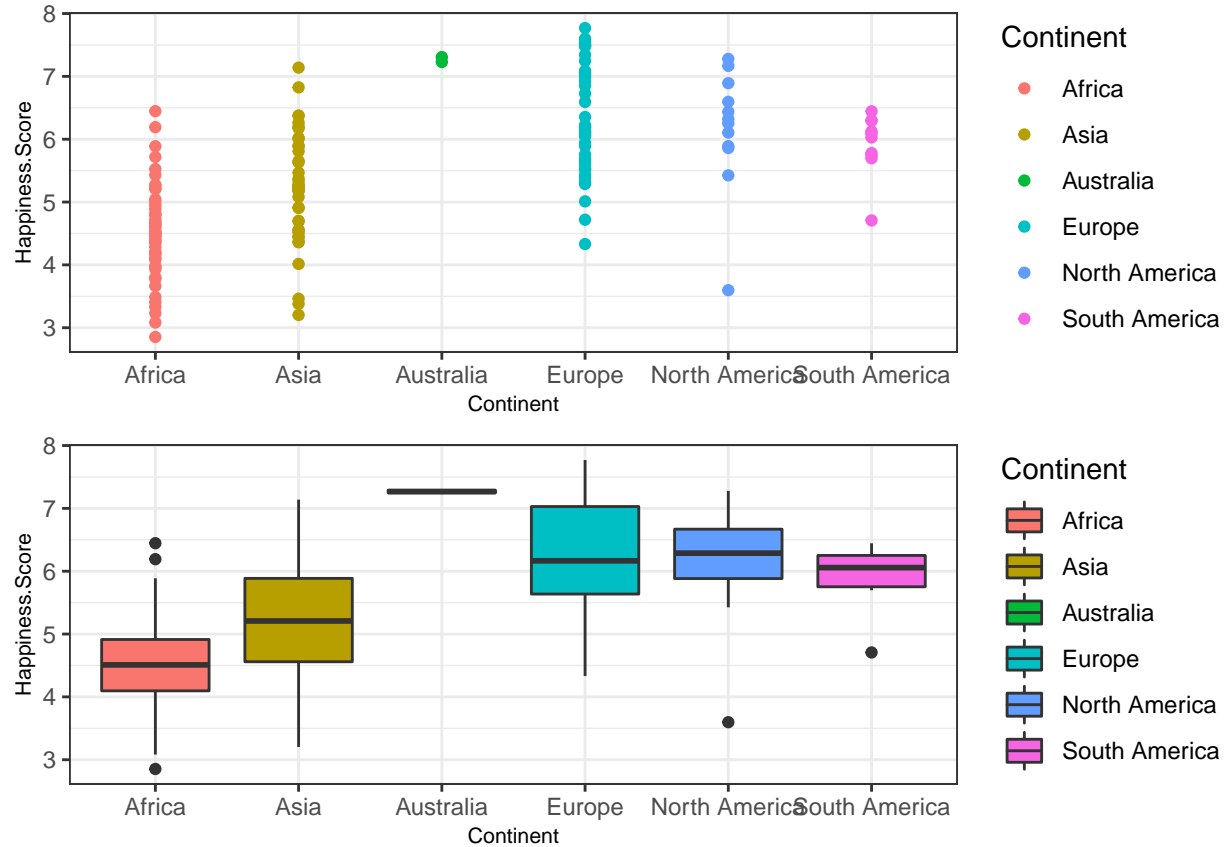
As we have seen before, Australia has the highest median happiness score. Europe, South America, and North America are in the second place regarding median happiness score. Asia has the lowest median after Africa. We can see the range of happiness score for different continents, and also the concentration of happiness score.

Scatter plot with regression line

Let's see the correlation between happiness score and the other seven factors in the happiness dataset for different continents by creating a scatter plot.

Scatter plot with regression line
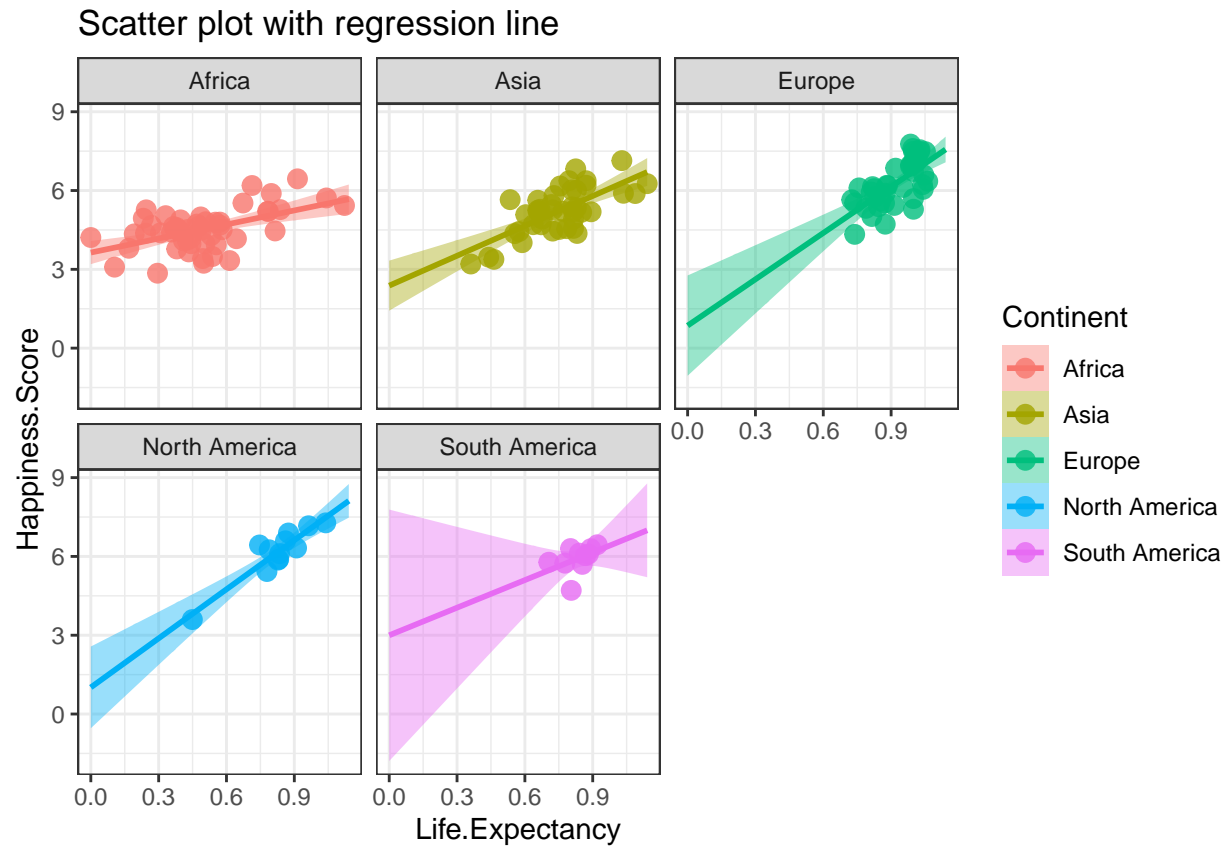
The correlation between life expectancy and happiness score in Europe, North America, and Asia is more significant than the other continents. Worth mentioning that we will not take Australia into account because there are just two countries in Australia and creating scatter plot with the regression line for this continent will not give us any insight.

Scatter plot with regression line
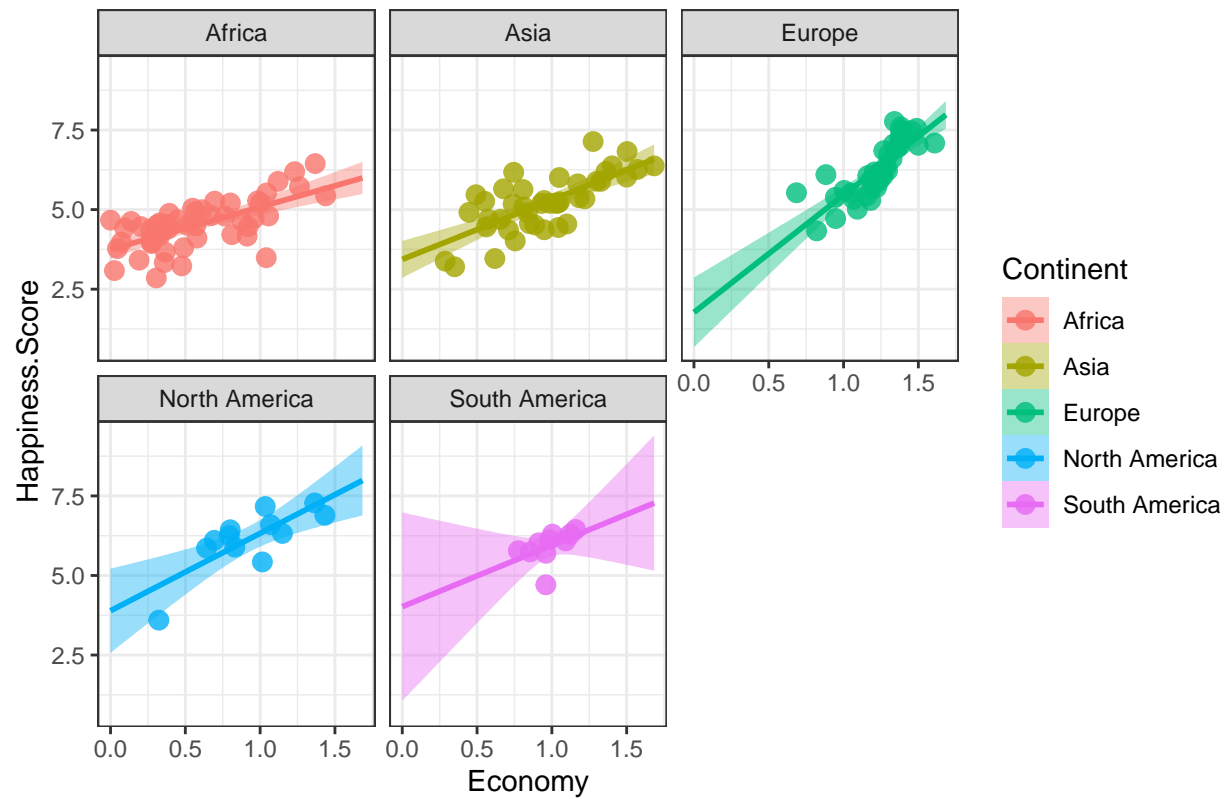
We can see pretty the same result here for the correlation between happiness score and economy. Africa has the lowest relationship in this regard.

Scatter plot with regression line
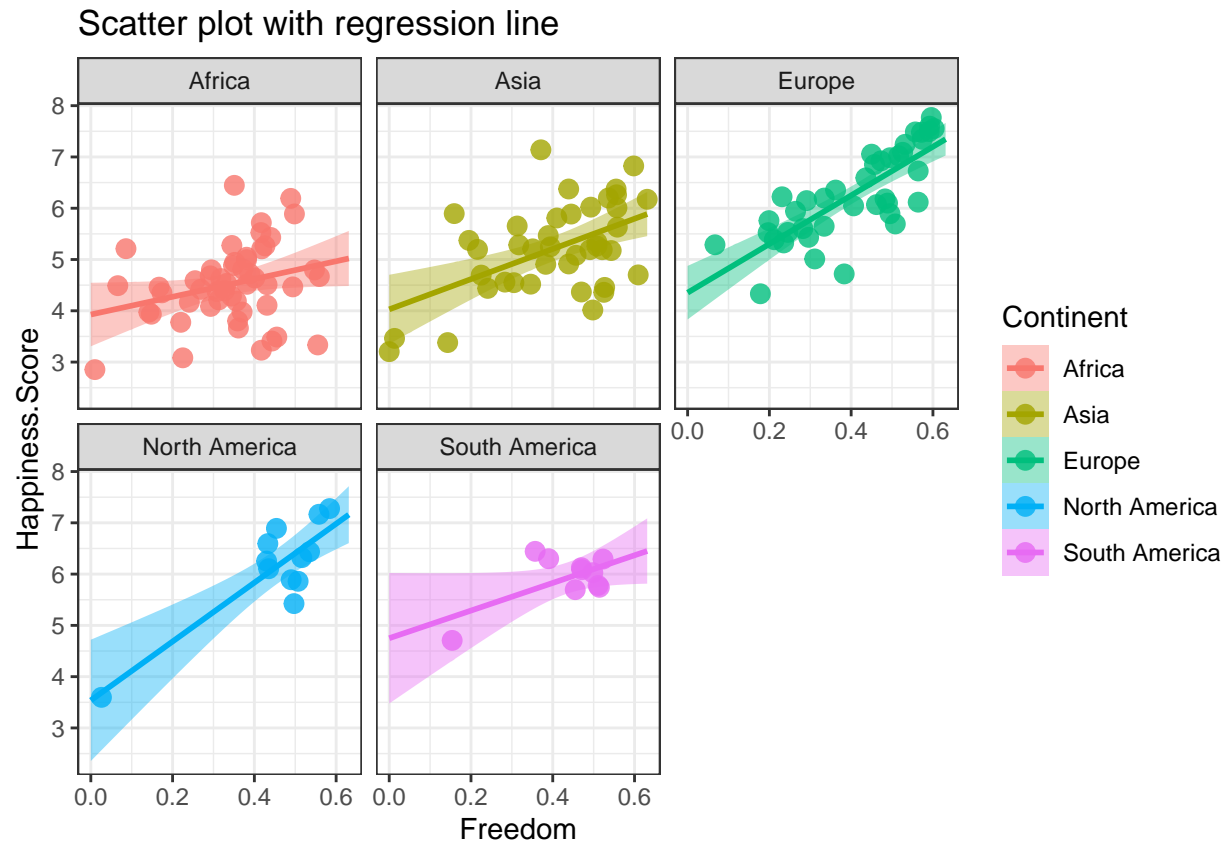
Freedom in Europe and North America is more correlated to happiness score than any other continents.

Scatter plot with regression line
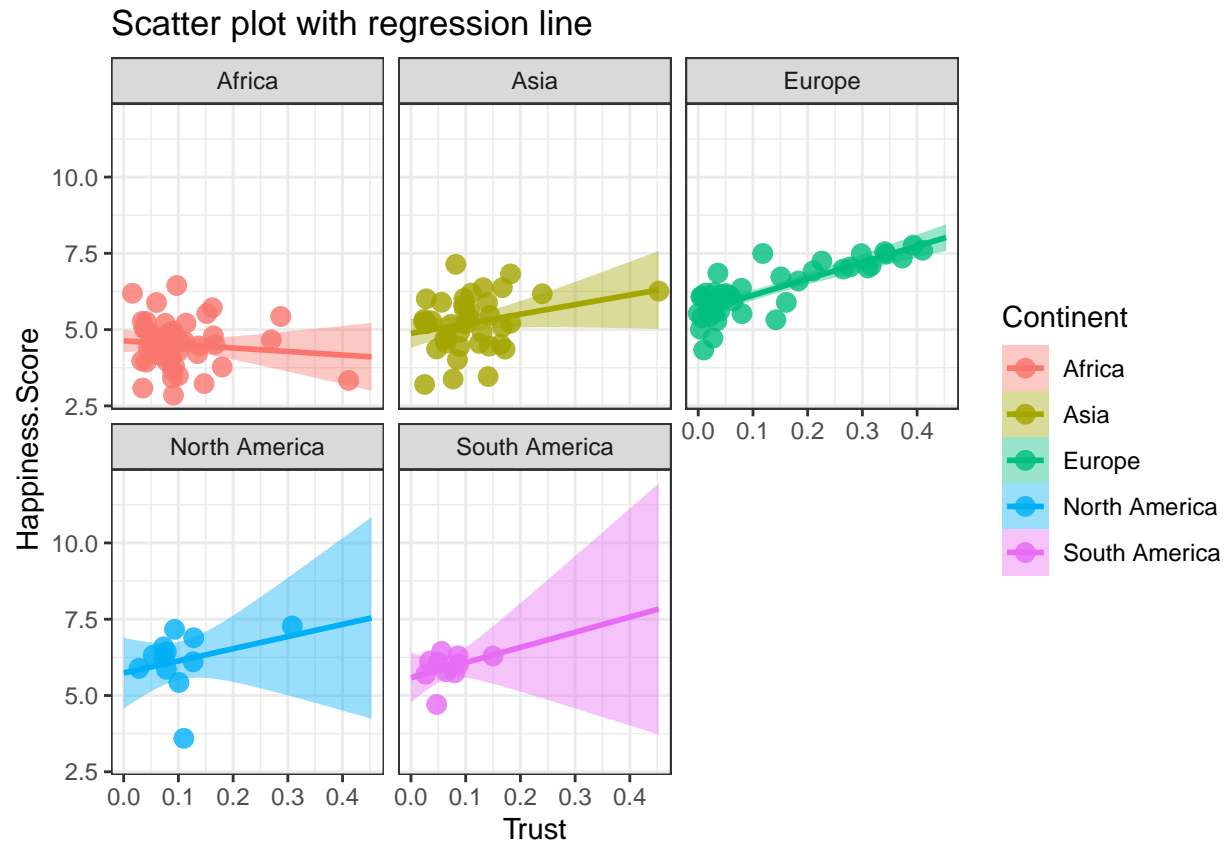
Approximately there is no correlation between trust and happiness score in Africa.

Scatter plot with regression line
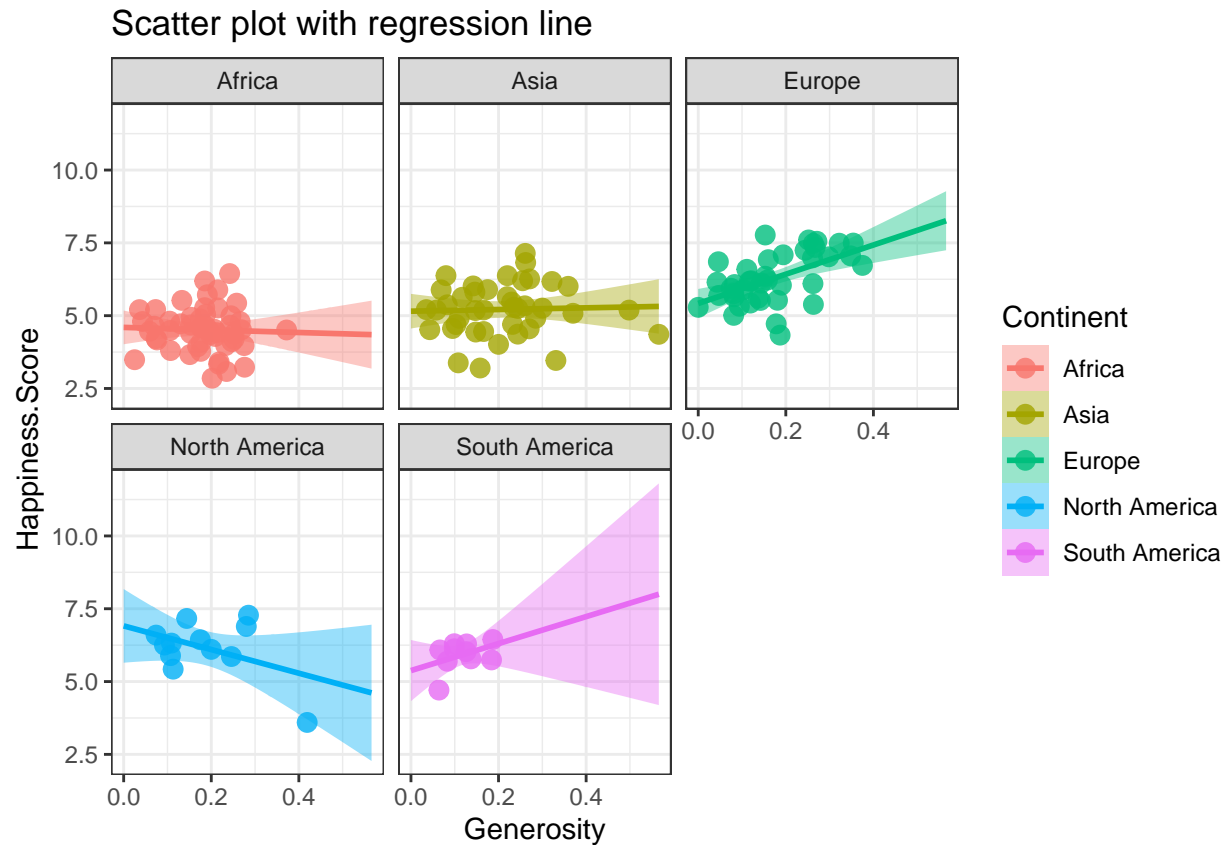
The regression line has a positive slope only for Europe and South America. For Asia the line is horizontal, and for Africa and North America the slope is negative.

Scatter plot colored by Continents

The following is just another way of seeing happiness score distribution on different continents when taking the correlation of happiness score with different variables into account.

R2 = 0.0057492 Intercept = 5.2433 Slope = 0.88606 P = 0.34682

R2 = 0.3212 Intercept = 3.6788 Slope = 4.4026 P = 1.2379e−14

R2 = 0.63025 Intercept = 3.3993 Slope = 2.2181 P = 4.3155e−35

R2 = 0.60822 Intercept = 2.8068 Slope = 3.5854 P = 3.7855e−33

R2 = 0.1487 Intercept = 4.9049 Slope = 4.5403 P = 6.654e−07
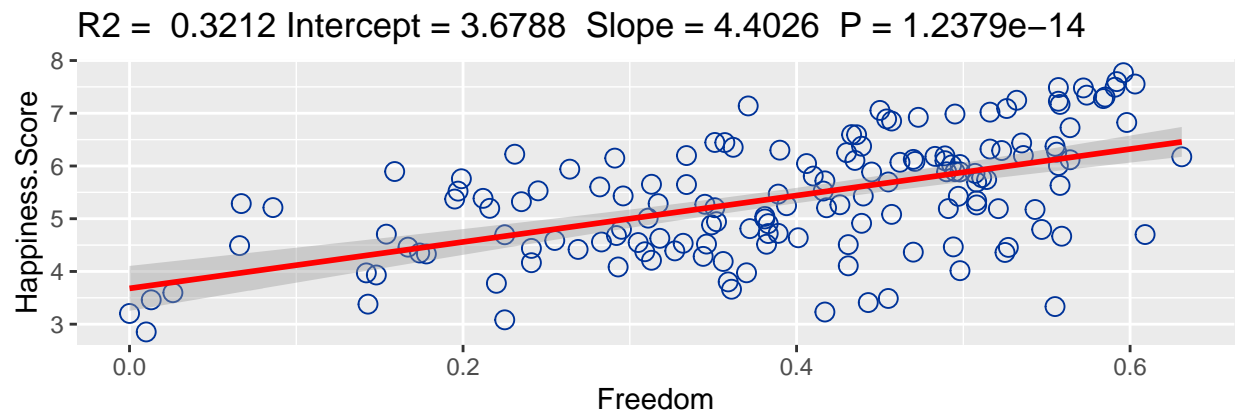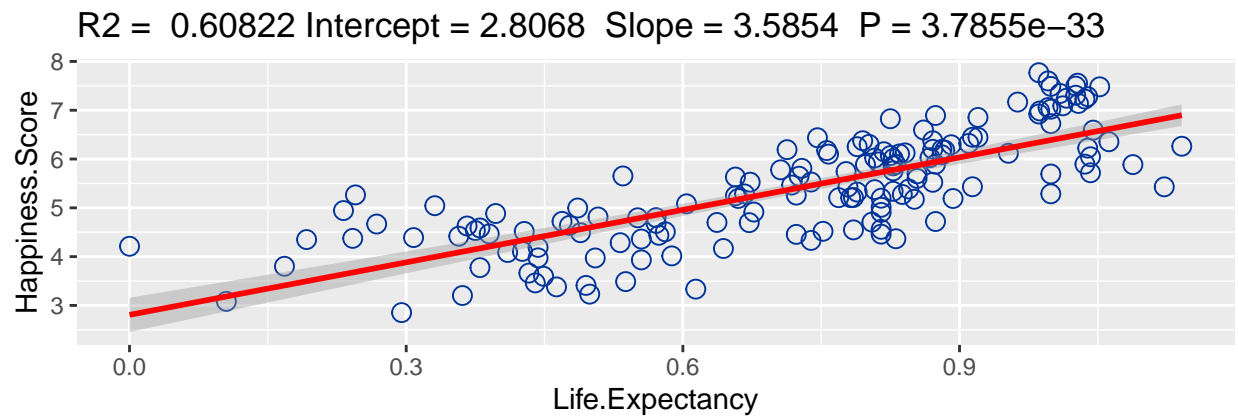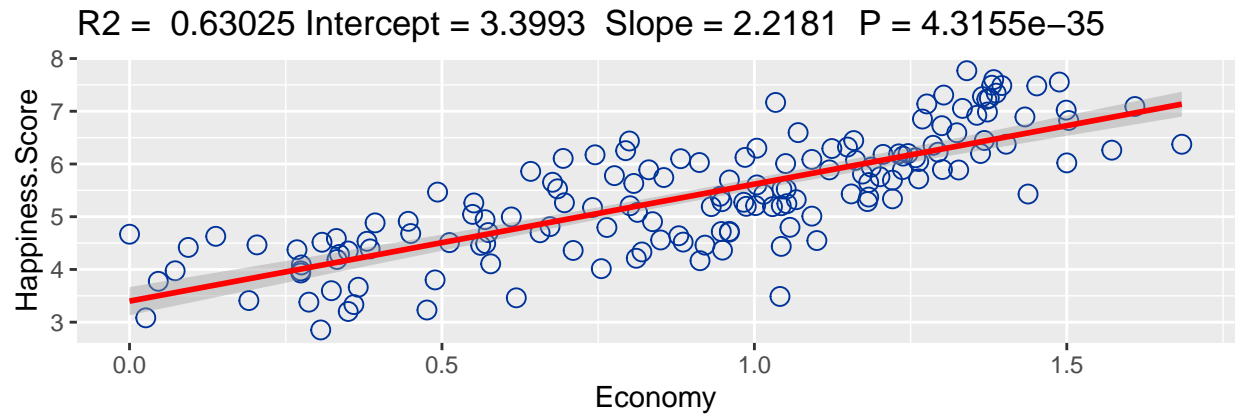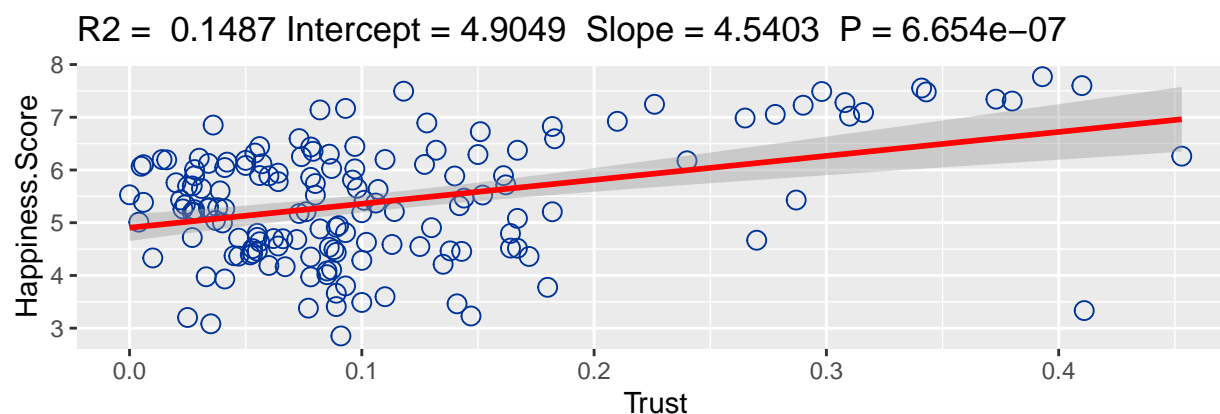
Visualizing data patterns may offer clues of intriguing routes to choose when performing analysis but it doesn't provide any noteworthy prediction capabilities. This is where machine learning techniques begin to excel. In the next section, we build several models to determine the best way to predict a happiness score.

Methods

Model 1: The Sum of Factors

The first model is seemingly the most obvious and was proposed on a discussion forum from the Kaggle website. It states that the "perfect" prediction model is to simply take the sum of all factors as the happiness score. This model is attempted with one caveat: a "standard dystopia score" was discovered in earlier happiness reports and was given the value 1.85. This value is added to our predicted scores as well because each factor is a ranking of how much *better* the country is than the standard dystopia. Note the use Root Mean Square Error (RMSE) as a success indicator. This choice is further explained in the results section.

```
##   Overall.rank Country.or.region Score pred_score      RMSE
## 1            1           Finland 7.769      6.905 0.5280065
## 2            2           Denmark 7.600      7.056 0.5280065
## 3            3            Norway 7.554      7.163 0.5280065
## 4            4           Iceland 7.494      6.943 0.5280065
## 5            5       Netherlands 7.488      6.944 0.5280065
```

All predicted scores shows the same RMSE! In addition to the dystopian standard score, previous Happiness Report data sets also have a "dystopian residual" that contributes to the happiness score. Since the residual

is not shown or described in this data set, it is deemed inappropriate to introduce it ourselved into the model. Although, without this value, the data set does seem incomplete. In spite of its absence, the dystopian residual column is calculated and a snippet of these are shown below. This is calculated under the assumption the summation model is perfect. Simply subtracting all factors and the standard dystopia value from the happiness score will yield the residual. There is no mention of the missing residual in any recent discussions to the 2018 or 2019 dataset, therefore it is omitted in subsequent modeling techniques. We can conclude that the sum of factors model is not perfectly accurate as portrayed in the discussion boards; it has a RMSE of 0.5280065.

```
##   Overall.rank Country.or.region Score pred_score      RMSE residual
## 1            1            Finland 7.769      6.905 0.5280065    0.864
## 2            2            Denmark 7.600      7.056 0.5280065    0.544
## 3            3             Norway 7.554      7.163 0.5280065    0.391
## 4            4            Iceland 7.494      6.943 0.5280065    0.551
## 5            5        Netherlands 7.488      6.944 0.5280065    0.544
```

Model 2: The 2019 GLM Model

Before our first linear regression model is applied, the data must be partitioned into a training and test set. This step is common when employing machine learning algorithms that require a check on the goodness of fit. It reduces the probability of overfitting to our training data at the expense of our prediction model. This was not completed for our sum of factors model because a model did not need to be trained, the equation was simply `sum(data$[factors])`.

The Happiness Report has a little over 150 country observations and six factors in which we will condition our model. Given the relatively low number of observations compared to the amount of factors, the model may have a tendency to overfit to the training data by overweighting unimportant variables. This is almost unavoidable when working with low volumes of data. The reality of regression is that you can *always* find a model that fits your training data exactly but which is typically useless for prediction. Keeping this in mind, an optimal training-test data split ratio, that is 70 training:30 test, 80 training:20 test, . . . , etc., is first determined.

The tested model in the partitioning is explored after this section. Plotting `rmses` versus `p` shows a slight pattern: when you increase the training data size, our RMSE decreases. Intuitively, this makes sense. The data is quite correlated as it has already been shown and the model is being allowed to work with more training data to make better predictions in the test set. From the plot below, the lowest RMSE is 0.3204118 with a ratio of 0.9:0.1.
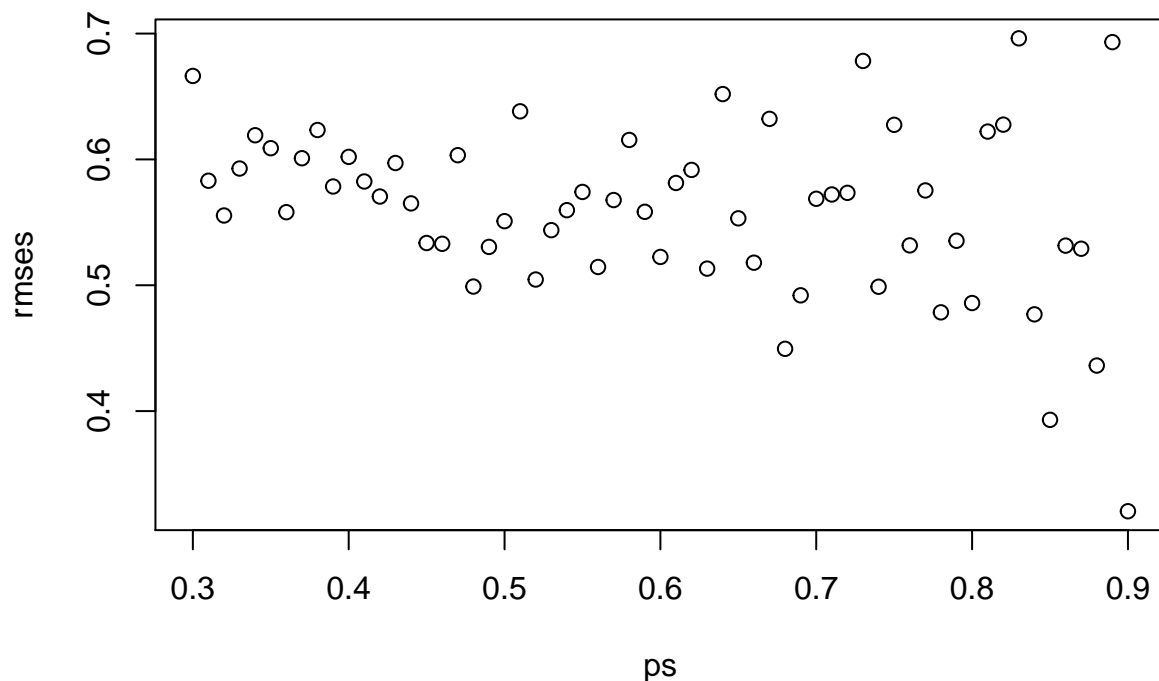
While useful in achieving a low RMSE, employing only 0.1 percent of our data to test does not leave much in terms of prediction. Ultimately, an arbitrary value of 0.70 is chosen because RMSE seems to become more sporadic after this value. Future models in the *methods* section use 0.70 as well. This ratio is also kept when the current data is supplemented with more data.

With our data partitioned 0.70:0.30, a generalized linear model is fitted using the `caret` package. `Score` is predicted using all six factors. The first five predicted scores are shown.

```
##   Overall.rank Country.or.region Score pred_score
## 1            1           Finland 7.769   6.958656
## 2            7            Sweden 7.343   6.986618
## 3           10           Austria 7.246   6.793396
## 4           12        Costa Rica 7.167   6.280958
## 5           17           Germany 6.985   6.720572
## 6           18           Belgium 6.923   6.587322
```

```
##    Overall.rank Country.or.region Score pred_score
## 39          140             India 4.015   4.918196
## 40          142           Comoros 3.973   3.930634
## 41          143        Madagascar 3.933   4.027239
## 42          144           Lesotho 3.802   4.350990
## 43          147             Haiti 3.597   3.819518
## 44          156       South Sudan 2.853   3.272902
```

The top five and bottom five observations are shown above compared with the actual score. The `results` data frame is plotted below with a line of best fit in blue and a reference line in red at `y = x`. If the model

was to work perfectly, the line of best fit would follow the reference line because each predicted score would be equal to the score.
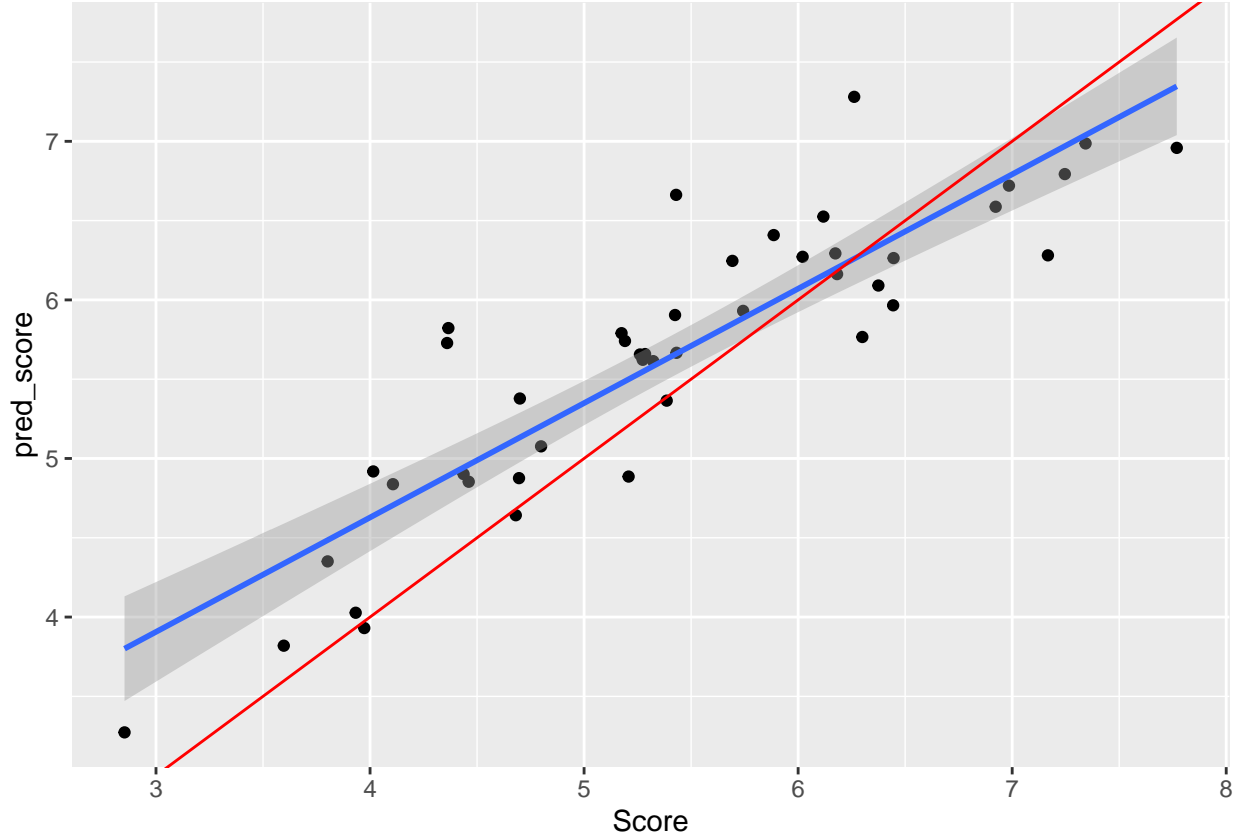
Even though results have been shown, it is worth mentioning the use of RMSE instead of other success indicators. RMSE suggests how close (or far) your predicted values are from the actual data you are attempting to model. The use of a success measure for this model, and others in the methods section, is to understand the accuracy and precision of the model's predictions. For this reason, RMSE is used as a success metric over alternatives. The RMSE of this model is 0.5718812. The coefficients of the fitted model are shown below for completeness:

```
##                 (Intercept)            GDP.per.capita
##                   1.8483663                 0.8230753
##              Social.support    Healthy.life.expectancy
##                   1.0155070                 1.0725756
## Freedom.to.make.life.choices                Generosity
##                   1.5864416                 0.9939111
##    Perceptions.of.corruption
##                   0.6122442
```

The following equation is the final model equation. Using these coefficients and the following notation predicted score $= \hat{y}$, GDP per capita score $= x_{GDP}$, Social Support score $= x_{SS}$, Life Expectancy score $= x_{HEA}$, Freedom score $= x_{FRE}$, Generosity score $= x_{GEN}$, Truth score $= x_{TRU}$.

$$\hat{y} = 1.848 + 0.823x_{GDP} + 1.016x_{SS} + 1.073x_{HEA} + 1.586x_{FRE} + 0.994x_{GEN} + 0.612x_{TRU} \qquad (1)$$

Removing Generosity

In an attempt to improve the model, we can remove the generosity component because early evaluations pointed to it being the least correlated factor. The previously partitioned data (`p = 0.70`) is used in this model as well.
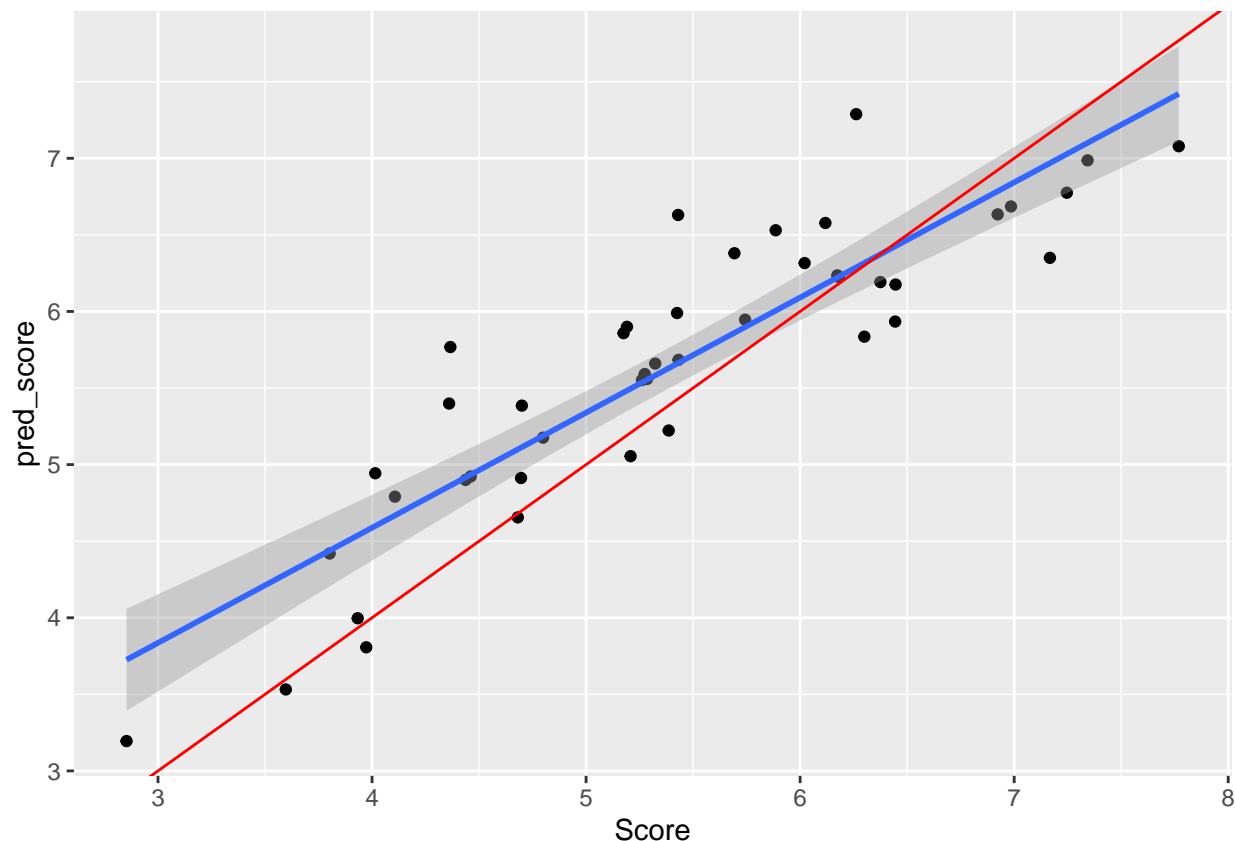
This model yields a RMSE of 0.5588842. The model coefficients are shown below along with the model equation following the same format as before.

```
##                   (Intercept)              GDP.per.capita
##                     1.9755643                   0.7931440
##                Social.support      Healthy.life.expectancy
##                     0.9743670                   1.0938625
## Freedom.to.make.life.choices     Perceptions.of.corruption
##                     1.8273152                   0.8300638
```

$$\hat{y} = 1.976 + 0.793x_{GDP} + 0.974x_{SS} + 1.094x_{HEA} + 1.827x_{FRE} + 0.83x_{TRU} \tag{2}$$

Model 3: The 2018/19 GLM Model

In an ideal world, data sets would be infintely large and models can be fitted with as much precision as your CPU can handle. The next model uses data from the 2018 and 2019 surveys. While not exactly an infinite data set, doubling the size of our train and test sets do allow for more training and testing. Combining data from old surveys is appropriate in this scenario because these two years follow the same grading scheme,

specifically they both do not have the dystopian residuals mentioned above. The first step is to reproduce new train and test sets, then run a model including all factors.

```
## Rows: 312
## Columns: 7
## $ Score                     <dbl> 7.769, 7.600, 7.554, 7.494, 7.488, 7.4...
## $ GDP.per.capita            <dbl> 1.340, 1.383, 1.488, 1.380, 1.396, 1.4...
## $ Social.support            <dbl> 1.587, 1.573, 1.582, 1.624, 1.522, 1.5...
## $ Healthy.life.expectancy   <dbl> 0.986, 0.996, 1.028, 1.026, 0.999, 1.0...
## $ Freedom.to.make.life.choices <dbl> 0.596, 0.592, 0.603, 0.591, 0.557, 0.5...
## $ Generosity                <dbl> 0.153, 0.252, 0.271, 0.354, 0.322, 0.2...
## $ Perceptions.of.corruption <dbl> 0.393, 0.410, 0.341, 0.118, 0.298, 0.3...
```

Columns with unused data, such as country name and rank, were removed before merging. NAs are introduced by coercion when converting the corruption column of the 2018 data set from factor to numeric. These are changed to zeros and represent unfound data. A glimpse of the 2018/2019 data set is outputted above.
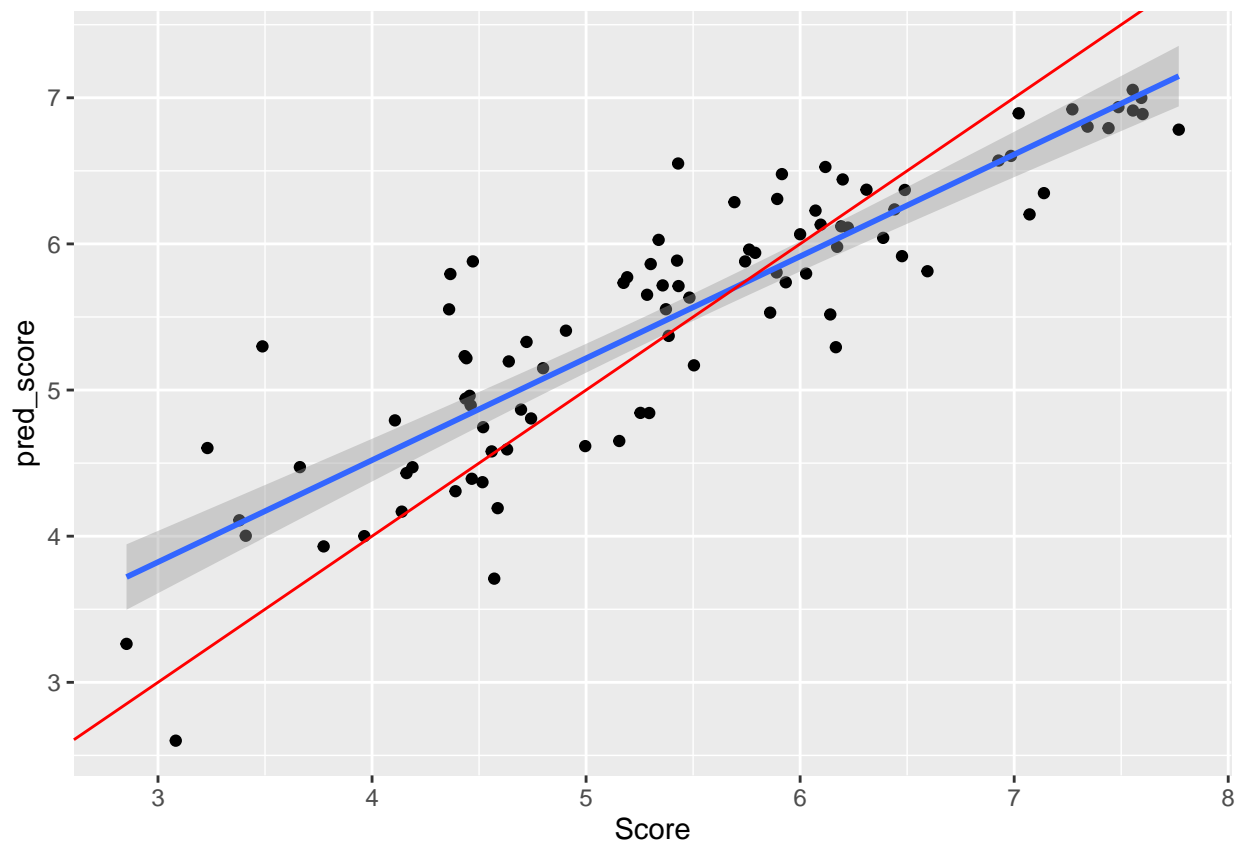
We see a better fitting model here, especially around the median, where the confidence bands are small. This model yields a RMSE of 0.5747943. The model coefficients are shown below along with the model equation following the same notation as before.

```
##              (Intercept)             GDP.per.capita
##               1.95281607                 0.95571540
##            Social.support    Healthy.life.expectancy
##               0.97041194                 0.96817249
```

32

```
## Freedom.to.make.life.choices                Generosity
##                 1.52136362                   0.75035656
##      Perceptions.of.corruption
##                 0.08183693
```

$$\hat{y} = 1.953 + 0.956x_{GDP} + 0.97x_{SS} + 0.968x_{HEA} + 1.521x_{FRE} + 0.75x_{GEN} + 0.082x_{TRU} \qquad (3)$$
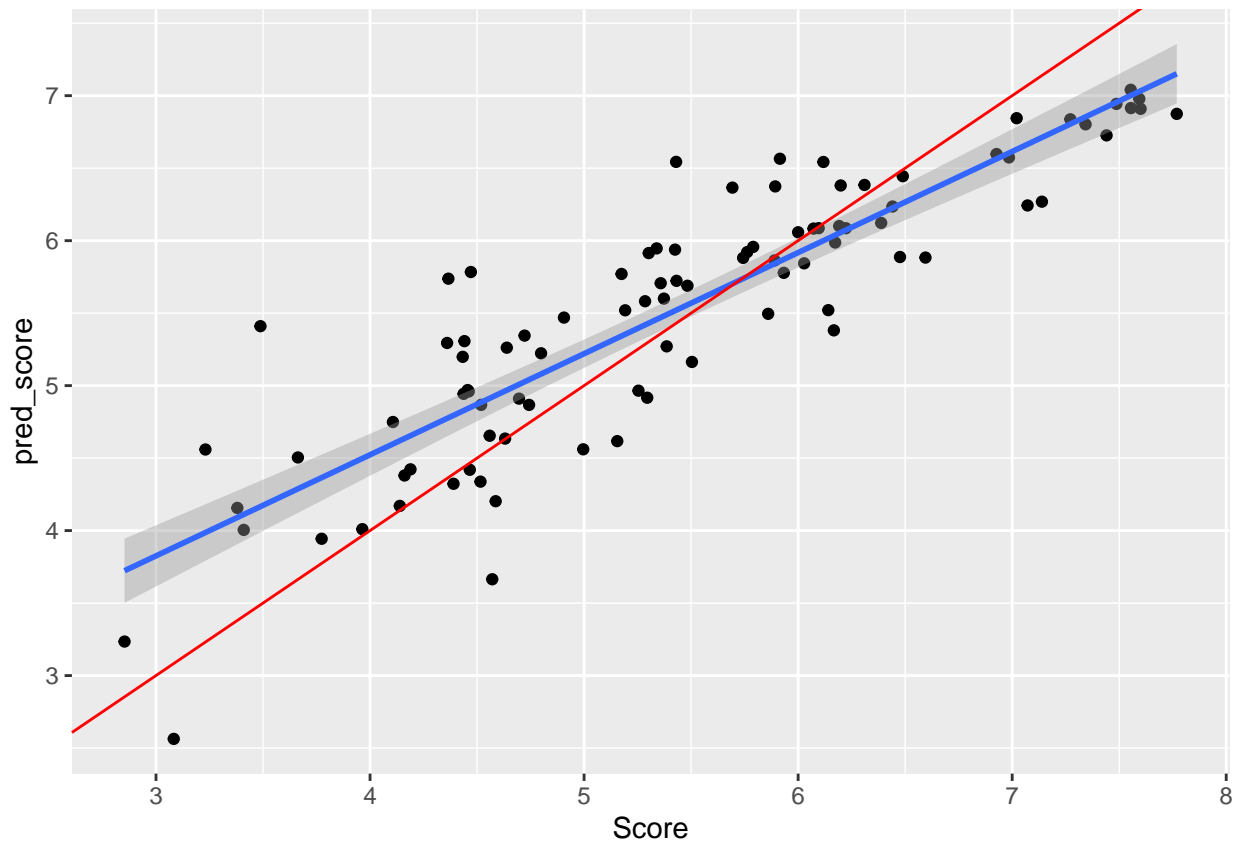
Removing Generosity

For curiosity's sake, the previous model may be improved by removing the somewhat uncorrelated generosity factor seen in earlier sections. This is completed with the 2018-2019 data set, `full_data`.

We see a better fitting model here, especially around the median, where the confidence bands are small. This model yields a RMSE of 0.5716972. The model coefficients are shown below along with the model equation following the same notation as before.

```
##                 (Intercept)                 GDP.per.capita
##                   2.0640555                      0.9153533
##              Social.support          Healthy.life.expectancy
##                   0.9602797                      0.9904659
## Freedom.to.make.life.choices      Perceptions.of.corruption
##                   1.5991209                      0.3312910
```

$$\hat{y} = 2.064 + 0.915x_{GDP} + 0.96x_{SS} + 0.99x_{HEA} + 1.599x_{FRE} + 0.331x_{TRU} \qquad (4)$$

Results

The results of our five models are shown in the table below. It is clearly shown that the best model in terms of RMSE is the sum of factors model. Even though the data seemed incomplete with dystopian residuals, taking the sum of the factor and the standard dystopian scores yield the closest score predictions.

Method	RMSE
Sum of Factors Model	0.528
GLM Model 2019	0.572
GLM Model 2018/2019	0.575
GLM Model - No Generosity 2019	0.559
GLM Model - No Generosity 2018/2019	0.572

For completeness, it may be useful to check the summation model with the full 2018-2019 data set. This is shown below and reported in the final table.

It is clearly seen that the best model is the summation model. Interestingly the GLM models do not improve when the data set size doubles. This may be due to overfitting in smaller data size that does not occur in the larger data.

Method	RMSE
Sum of Factors Model	0.528
Sum of Factors Model 2018/2019	0.526
GLM Model 2019	0.572
GLM Model 2018/2019	0.575
GLM Model - No Generosity 2019	0.559
GLM Model - No Generosity 2018/2019	0.572

Conclusion

Machine learning algorithms have become a paramount tool in computer science and analysis. This report explores this use with regression analysis on a dataset of national happiness. Results are found that are rather unexpected: a glaring error in the data did not impede the simplest model, increasing training data set size did not decrease the error, and reducing the effectiveness of uncorrelated variables did not decrease the error.

This report felt limited in the amount of data being trained and tested. If more consistent data was available, it is my belief that the summation model would out perform the GLM models. The omitted residuals left a hole in the data that may have provided more insight into the validity of the summation model. The discussion boards were helpful in this respect by providing insight on the purpose of the data. Further work on a project such as this would include calculating and using dystopian residuals for data sets where they were lost, combining previous years reports, and comparing more adaptive models on the large data sets.