What Is The Happiest Country In The World?

Jane Farris

12/01/2020

Contents

Overview

This project is part of the HarvardX Professional Certificate in Data Science capstone course. It consists of three files: a report in the form of an Rmd file, a report in the form of a PDF document (knit from the Rmd file), and the R script that generates the predicted movie ratings. This report contains the objective, data cleaning, exploratory analysis, data visualization, modeling, results and concluding remarks of the project. The dataset generated in this report is available on the official World Happiness Report website as an xls file.

Introduction

The World Happiness Report is a well recognized survey of the state of global happiness developed using the data from the Gallup World Poll (GWP). The report ranks countries by their happiness scores based on answers to the main life evaluation question asked in the poll, "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest". The first World Happiness Report was published in 2012 and has since continued to gain global recognition among governments and organizations as happiness indicators are used to establish policy making decisions. In fact, many professionals across various disciplines argue that the well-being or overall happiness of a country can be used as a measurement to assess the progress of a nation. This project analyzes the data from the 14 year period 2005-2018, including the variations in happiness between different countries, regions and years.

Objective

The aim of this project is to predict the subjective well-being or happiness score of a given country (ranging from 0 to 10), using the World Happiness Report data from 2005-2018. Analysis of various social, economic and political data points is carried out in order to explore their relationship to a countries overall well-being.

Additionally, we want to learn which countries or regions rank the highest in overall happiness and in each of the factors contributing to happiness such as GDP, healthy life expectancy, government corruption, social support, freedom and generosity. Finally we want to examine country scores and rankings over the 14 year period to see any major trends and changes.

This report contains 4 models developed to predict the happiness score of a given country, using 4 different machine learning techniques. Multivariate linear regression, decision (regression) trees, random forests and k-nearest neighbors analysis, using the original variables from the World Happiness Report were applied. The accuracy of each machine learning technique will be evaluated using Root Mean Square Error (RMSE) in order to see how far off the model predictions are from the observed country happiness scores reported in 2005-2018. The RMSE formula is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2}$$

The Dataset

The World Happiness Report of 2019, which includes data from 2005-2018 can be found and downloaded at: https://worldhappiness.report/ed/2019/.

The final dataset containing the years 2005-2018 of World Happiness Report data contains 1704 rows each corresponding to countries around the world and the following 12 columns:

Country: Name of the country.

Region: Region the country belongs to.

Score: The happiness score of a country; a metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest". The national average of the response to the question.

Log.GDP: the logarithm of a country's GDP per capita in the respective year.

Social.support: the national average of the binary responses (either 0 or 1) to the GWP 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?"; having someone to count on in times of trouble.

Healthy.life.expectancy: the average number of years a person can expect to live in full health, without disabling illnesses or injuries. Based on data reported from the World Health Orga- nization (WHO).

Freedom: the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?".

Perceptions.of.corruption: the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"; the overall perception is just the average of the two binary responses.

Generosity: the national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.

Year: The year the World Happiness Report data was published.

Positive.affect: the average of three positive affect measures in GWP: happiness, laugh and enjoyment.

Negative.affect: the average of three negative affect measures in GWP: worry, sadness and anger.

Methods & Analysis

This section of the report contains the data exploration, visualization and modeling techniques used.

Data Exploration

First the data was manipulated and summarized in order to initially analyze its characteristics and prepare it for testing and model building. We can see we have 1704 observations of 12 variables. Each row represents a country's overall happiness score in a given year (ranging from 2005 to 2018) and includes other variables corresponding to different economic, political and social topics. There are 165 different countries represented in our dataset, each categorized into one of 10 different regions: Southern Asia, Eastern Asia, Southeastern Asia, Central and Eastern Europe, Western Europe, Middle East and North Africa, Sub-Saharan Africa, Latin America and Caribbean, North America and Australia and New Zealand.

```
#Dataset size
dim(happiness_data)
```

```
## [1] 1704   12
```

```
#How many countries are in our dataset
length(unique(happiness_data$Country))
```

```
## [1] 165
```

```
#How many regions are in our dataset
length(unique(happiness_data$Region))
```

```
## [1] 10
```

```
#Range of world happiness reports
min(happiness_data$Year)
```

```
## [1] 2005
```

```
max(happiness_data$Year)
```

```
## [1] 2018
```

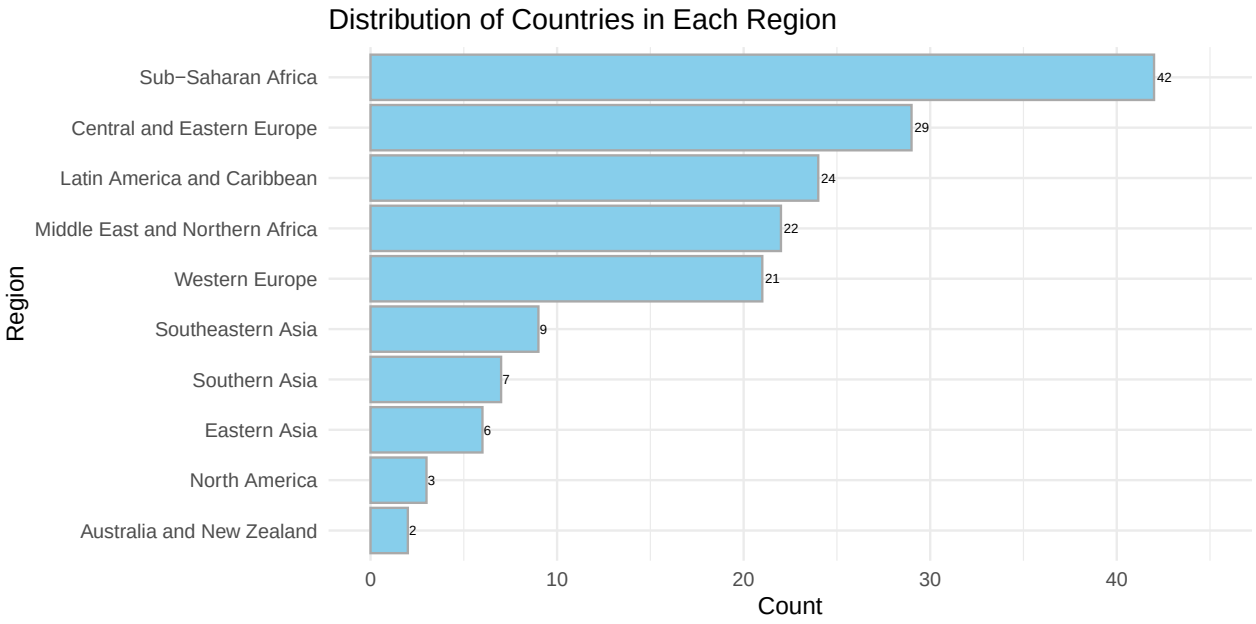
Average World Happiness

We can calculate the average overall world happiness between 2005-2018 to see if there are any major trends. We can see that 2005 had the highest overall well-being score by a lot, while between 2006-2018 the score decreased and fluctuated around 5.4. This could be caused by many different world events/trends such as the rise of social media, climate change, oil prices, election results and so on. It is virtually impossible to conclude the exact source of the decline in overall world happiness, thus we can only speculate based on world events.

Year	Average Happiness Score
2005	6.446164
2006	5.196935
2007	5.418291
2008	5.418554
2009	5.457640
2010	5.496782
2011	5.424088
2012	5.443751
2013	5.394568
2014	5.389023
2015	5.404037
2016	5.399414
2017	5.460421
2018	5.502134

Distribution of Countries in Each Region

Next, we can further explore the breakdown of countries and regions included in our dataset. We can see that we have 10 different regions, each representing a different amount of countries. For instance, Sub-Saharan Africa and Central/Eastern Europe contain the most amount of countries with 42 and 29 respectively. Conversely, the region Australia and New Zealand only contains 2 countries (Australia and New Zealand).



Happiest Countries

Below is a list of the top 10 happiest countries in each year from 2005 to 2018. We can see there is a lot of similarity between the lists over the years, with countries like Denmark, Finland, Netherlands, Norway, Canada, Australia etc. represented in almost every year. In fact, Denmark and Finland both hold the top spot quite frequently, with Denmark being the happiest country 7 times and Finland 4 times in the 14 year period. Further, it's important to note that most of the countries that made the list are from the Western European region and no country from Sub-Saharan Africa is represented. This is interesting because the Sub-Saharan African region contains the most amount of countries by a lot, yet none of the have made the top happiest countries list in the 14 year period.

Rank	2005	2006	2007	2008	2009	2010	2011
1	Denmark	Finland	Denmark	Denmark	Denmark	Denmark	Denmark
2	Netherlands	Switzerland	New Zealand	Finland	Costa Rica	Canada	Netherlands
3	Canada	Norway	United States	Norway	Switzerland	Netherlands	Austria
4	Sweden	New Zealand	Canada	Netherlands	Canada	Sweden	Israel
5	Australia	United States	Netherlands	Ireland	Israel	Venezuela	Canada
6	Belgium	Israel	Costa Rica	Sweden	Sweden	Australia	Australia
7	Venezuela	Ireland	Australia	Canada	Venezuela	Finland	Sweden
8	Spain	Austria	Saudi Arabia	New Zealand	United States	Israel	Finland
9	France	Costa Rica	Sweden	Spain	Ireland	Panama	Panama
10	Saudi Arabia	United Arab Emirates	Belgium	United States	Panama	Austria	Costa Rica

2012	2013	2014	2015	2016	2017	2018
Switzerland	Canada	Denmark	Norway	Finland	Finland	Finland
Norway	Denmark	Switzerland	Switzerland	Norway	Denmark	Denmark
Iceland	Iceland	Norway	Denmark	Denmark	Norway	Switzerland
Sweden	Austria	Israel	Iceland	Netherlands	Iceland	Netherlands
Denmark	Finland	Finland	Finland	Iceland	Switzerland	Norway
Netherlands	Mexico	Netherlands	New Zealand	Switzerland	Netherlands	Austria
Finland	Sweden	New Zealand	Canada	Sweden	Canada	Sweden
Canada	Netherlands	Canada	Netherlands	Australia	Israel	New Zealand
Austria	Australia	Australia	Australia	Canada	New Zealand	Luxembourg
Mexico	Israel	Costa Rica	Sweden	New Zealand	Austria	United Kingdom

Least Happy Countries

Below is a list of the 10 least happy countries in each year from 2005 to 2018. We can see that similar to the top 10 happiest countries table, there is a lot of repetition of countries over the years. However, now there a very few countries from the Western Europe region. Instead, most of the countries over the years belong to the Sub-Saharan Africa, Southern Asia or Middle East and Northern Africa regions. This is likely due to the political climate, poverty and limited access to medical care, education and resources of most of the countries in these regions.

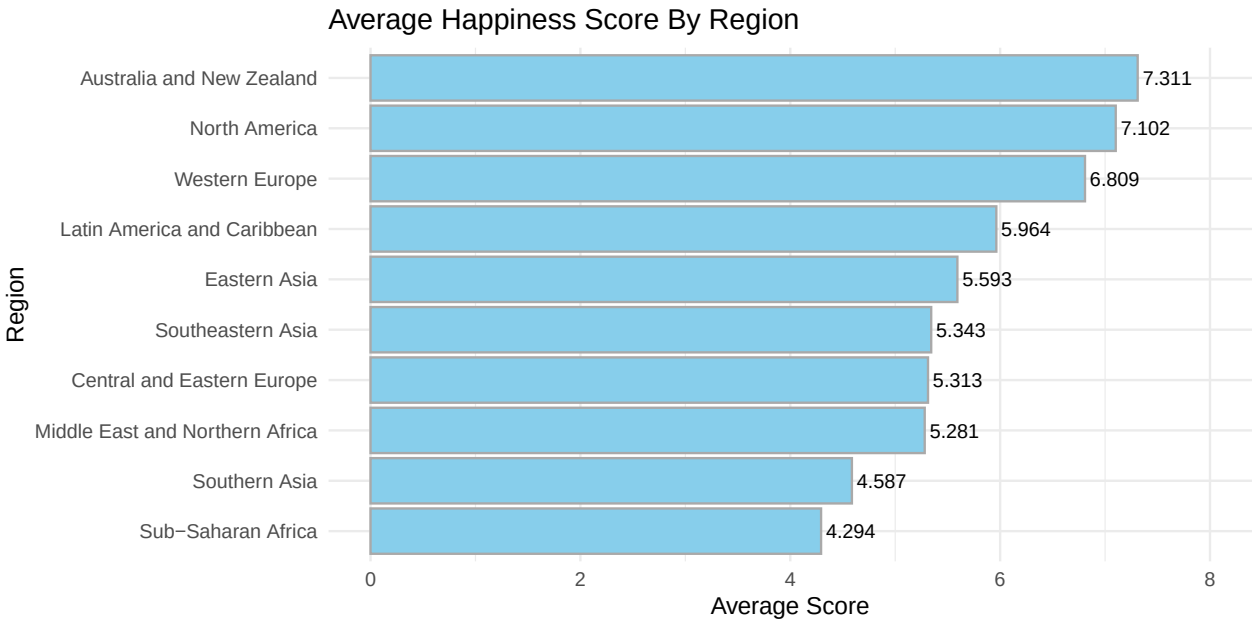
Rank	2005	2006	2007	2008	2009	2010	2011
1	Turkey	Togo	Zimbabwe	Togo	Tanzania	Tanzania	Togo
2	Romania	Benin	Sierra Leone	Sierra Leone	Comoros	Botswana	Botswana
3	Egypt	Chad	Liberia	Zimbabwe	Chad	Central African Republic	Central African Republic
4	Hungary	Cambodia	Georgia	Burundi	Burundi	Chad	Burundi
5	Pakistan	Sierra Leone	Bulgaria	Benin	Georgia	Mali	Yemen
6	Iran	Georgia	Zambia	Afghanistan	Mali	Haiti	Nepal
7	Lebanon	Uganda	Burkina Faso	Congo (Brazzaville)	Congo (Kinshasa)	Comoros	Afghanistan
8	Poland	Niger	Chad	Haiti	Rwanda	Bulgaria	Senegal
9	Greece	Haiti	Mauritania	Burkina Faso	Zimbabwe	Sri Lanka	Comoros
10	Jordan	Burkina Faso	Palestinian Territories	Kenya	Cambodia	Burkina Faso	Benin

2012	2013	2014	2015	2016	2017	2018
Syria	Syria	Togo	Liberia	Central African Republic	Afghanistan	Afghanistan
Benin	Burkina Faso	Burundi	Yemen	South Sudan	South Sudan	Yemen
Rwanda	Rwanda	Afghanistan	Syria	Tanzania	Rwanda	Malawi
Madagascar	Benin	Benin	Rwanda	Rwanda	Yemen	Tanzania
Guinea	Chad	Guinea	Guinea	Haiti	Tanzania	Botswana
Senegal	Egypt	Chad	Haiti	Liberia	Malawi	Rwanda
Afghanistan	Afghanistan	Burkina Faso	Madagascar	Malawi	Central African Republic	Haiti
Niger	Senegal	Tanzania	Benin	Botswana	Botswana	Zimbabwe
Cambodia	South Africa	Ivory Coast	Tanzania	Guinea	Zimbabwe	Burundi
Congo (Brazzaville)	Cambodia	Rwanda	Niger	Madagascar	Lesotho	India

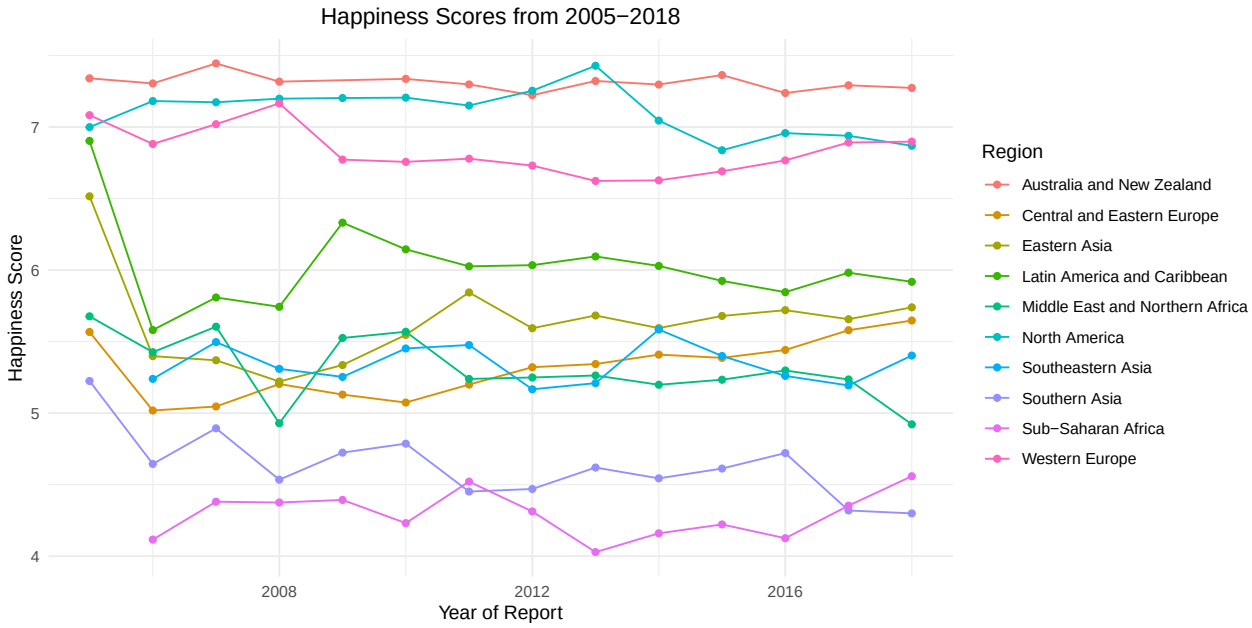
Data Visualiztion

Happiness By Region

Intuitively we know that certain regions will tend to have more overall happiness than others. We can see that the regions Australia and New Zealand, North America, Western Europe, Latin America and the Carribean and Eastern Asia all have happiness scores higher than the global average of 5.437. Again we see that countries belonging to Southern Asia and Sub-Saharan Africa have a significantly lower average happiness score in comparison to the rest of the world.

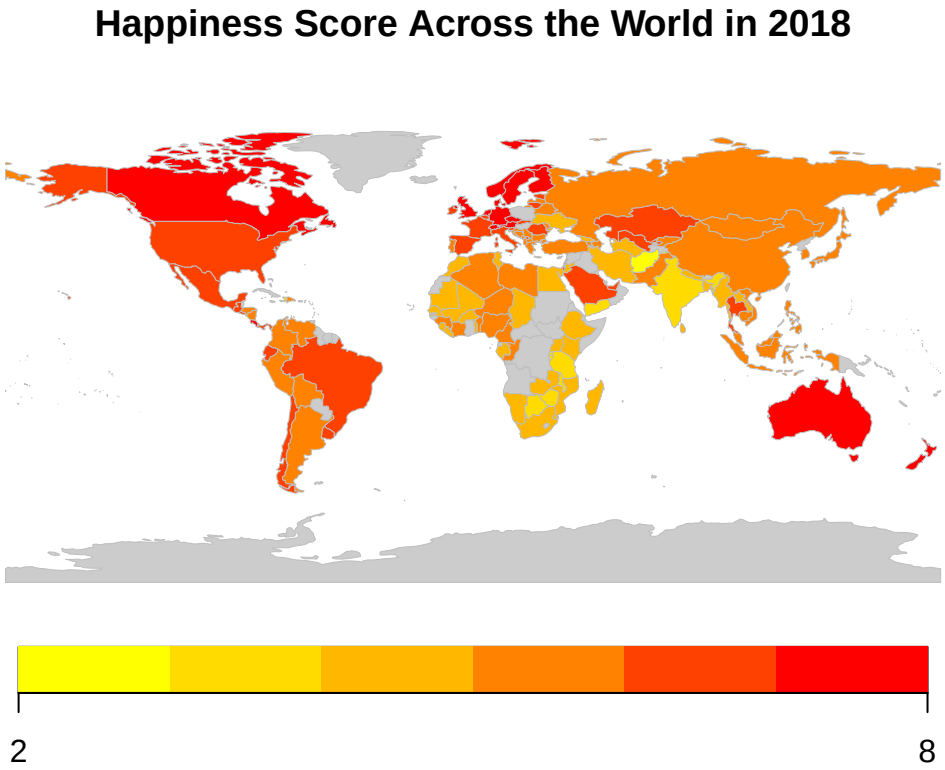


We can also explore the average happiness of each of the regions between 2005-2018 to see if there were any major changes. From the graph below we can see that Australia and New Zealand had the highest average scores for 12 of the 14 years, with North America taking the top spot in those 2 years. Conversely, Southern Asia and Sub-Saharan Africa continuously have the lowest average scores during the entire 14 year period. Central and Eastern European countries had a steady incline in average happiness between 2010 and 2018, after a sharp decline in 2006. Countries in Southeastern Asia, Eastern Asia, Middle East and North African and Latin American and the Caribbean regions on the other hand had a significant amount of fluctuation over the 14 years, however they all remained in the middle of the pack.



Happiness Across the Globe

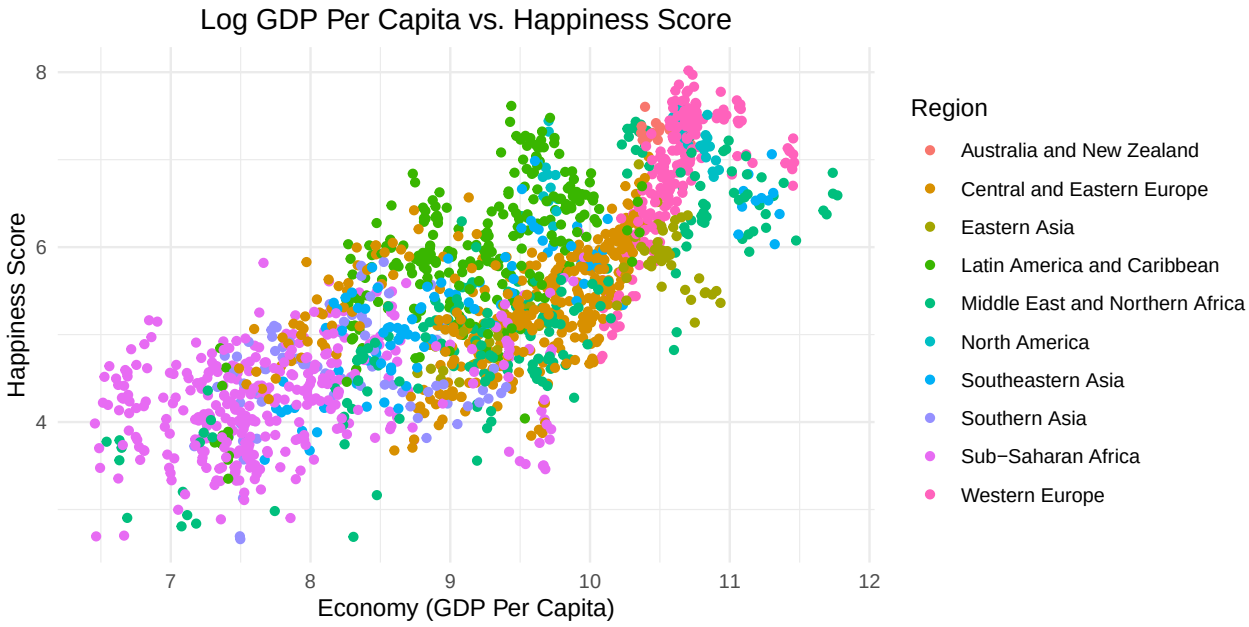
Below we can more clearly see the average happiness scores globally in 2018, which range from 2-8. This map supports the findings above that certain regions (such as Australia and New Zealand, North America and Western Europe) tend to have countries with higher overall happiness. Countries that are not included in our dataset are colored in light grey.



In order to build the most accurate model in predicting a country's happiness score, we must explore the relationship between our dependent variable and the features in our dataset. The following graphs examine each of the relationships:

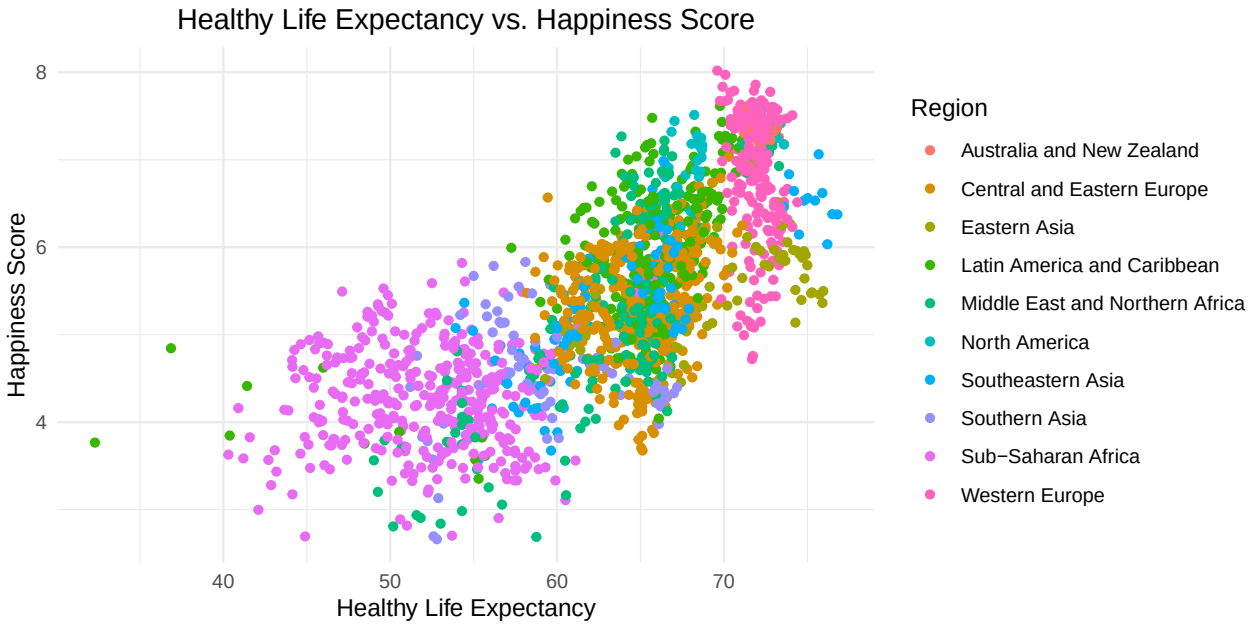
GDP vs. Happiness

First of all, we know intuitively that the GDP of a nation, the total monetary value of all the finished goods and services produced within a country's borders, is extremely important in predicting overall happiness. This is because GDP can be considered a measure of a country's economy, which when thriving leads to more jobs, trade and overall prosperity. This intuition is verified in the following graph of the relationship between the logarithm of a country's GDP and their well-being score. We can see that there is a very strong positive relationship, indicating that the higher a country's GDP is, ie. the larger the economy, the more likely a country is to have a higher happiness score. We can see that in general countries in the Sub-Saharan Africa region have a lower GDP and happiness score.



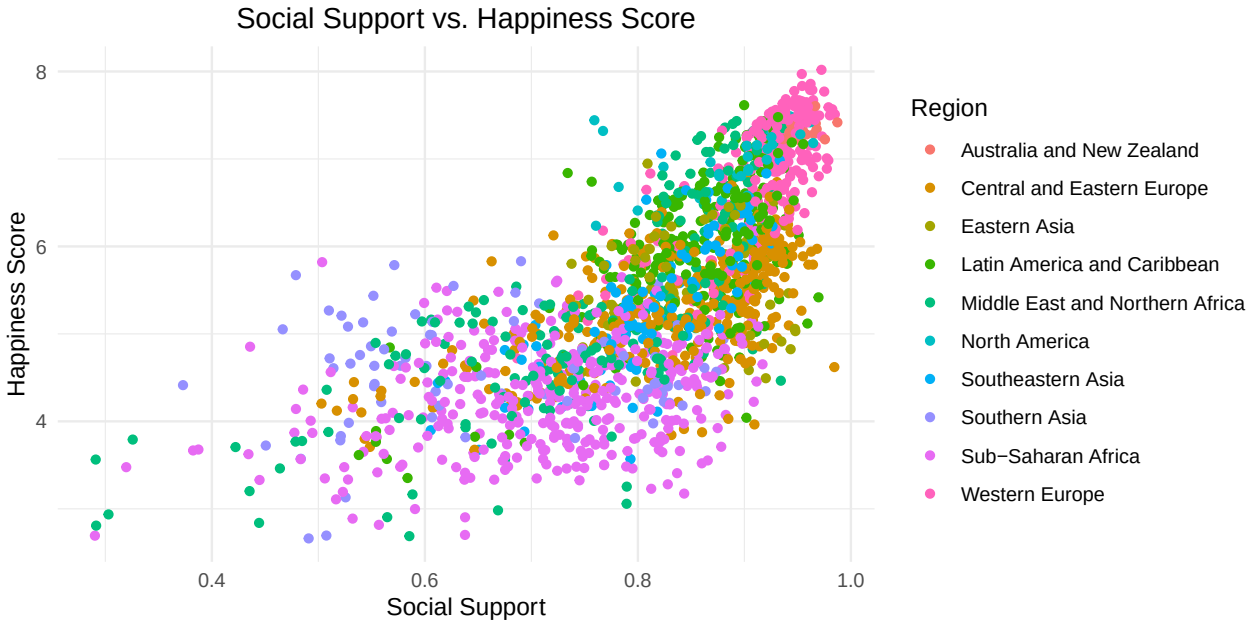
Healthy Life Expectancy vs. Happiness

Next, we can see that there is also a very strong positive relationship between healthy life expectancy and happiness. The average number of years that a person can expect to live in full health (not counting the years lived in less than full health due to disease and/or injury) is a very important indicator of overall happiness because it demonstrates a country's poverty level and medical capabilities/resources. We can see that countries in Western Europe and Eastern Asia have very high healthy life expectancies, with almost all of the countries in those regions having a HLE of above 70 years old.



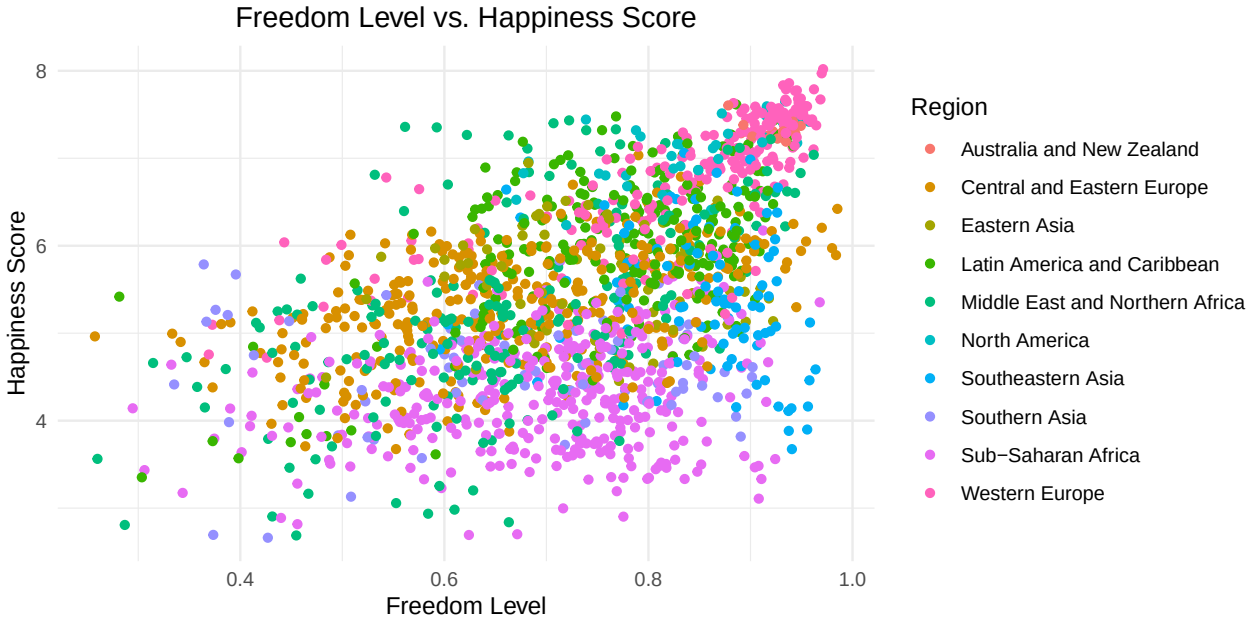
Social Support vs. Happiness

Social support, or having someone to count on in times of trouble, is another significant indicator of overall happiness. We can see in the graph below of the the relationship of the two variables that there is a strong positive relationship. This shows that the countries with people that feel they have a support system of friends and family tend to have higher well-being scores. We can see from the graph that the countries in the Western Europe region have the highest degree of social support and also some of the highest happiness scores.



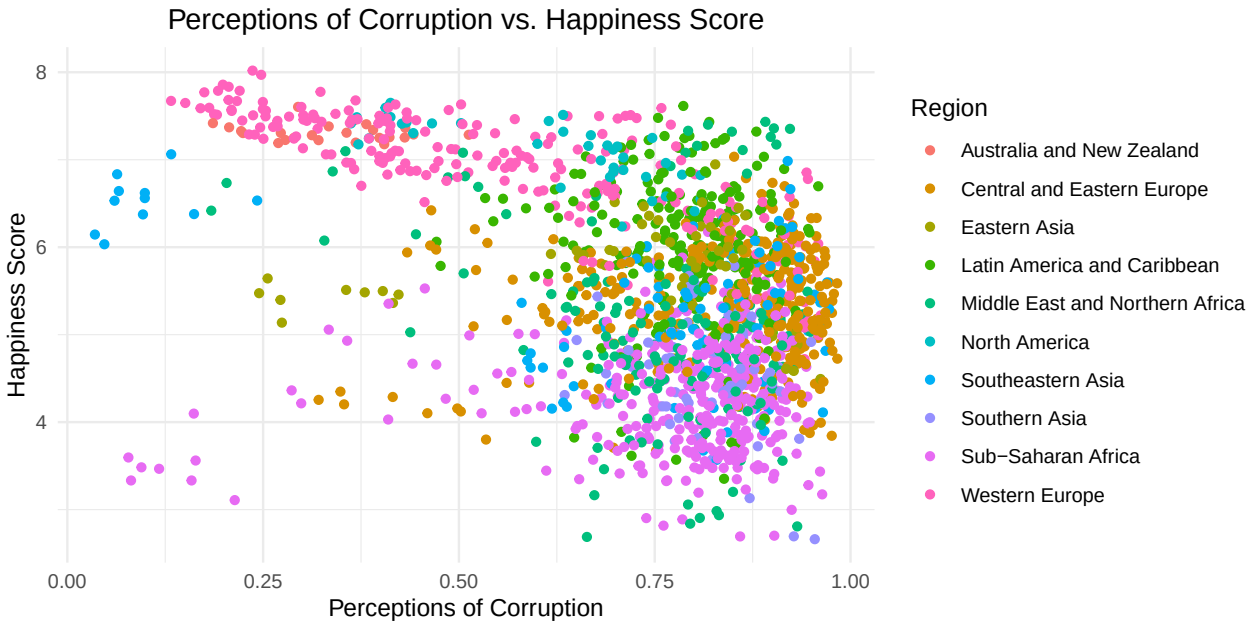
Freedom vs. Happiness

Looking at the graph below of the relationship between the freedom variable and happiness score, we can see that there is a positive relationship between the two, however it is slightly weaker. Again, countries in Western Europe tend to have the highest levels of freedom and also happiness scores. On the other hand, countries in the Middle East and North Africa, Sub-Saharan Africa and Southern Asia tend to have lower levels of freedom of life.



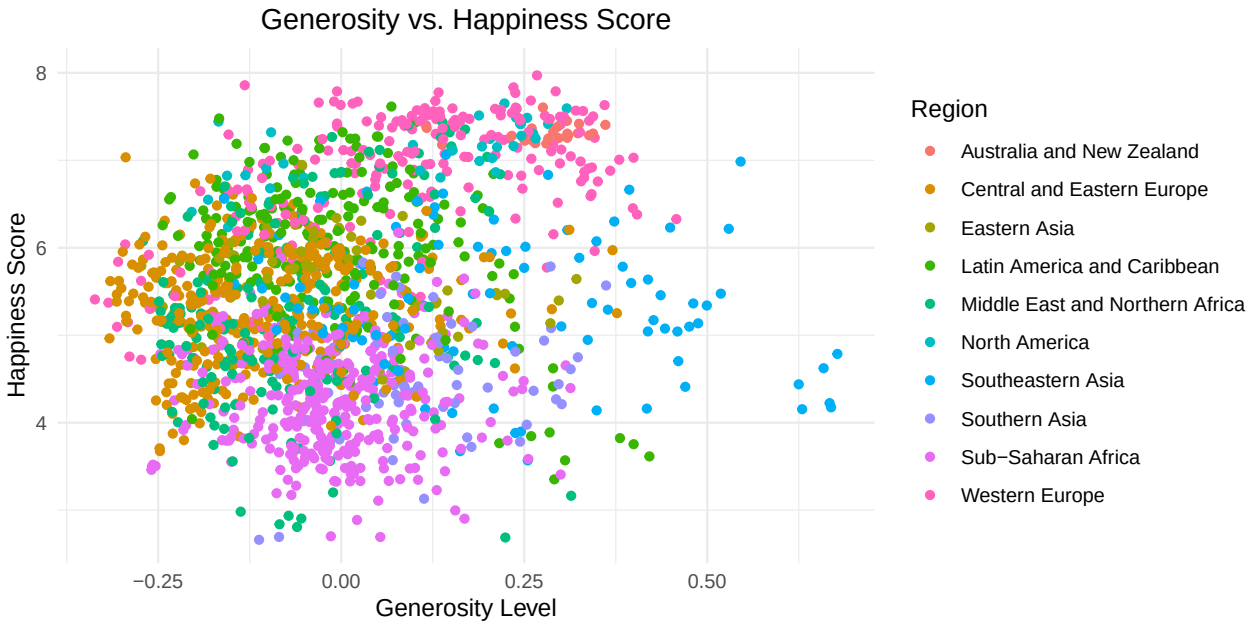
Perceptions of Corruption vs. Happiness

The graph of the perceptions of corruption, both within business and government, versus overall happiness is less clear than the previous relationships. It is apparent that countries in Central and Eastern Europe, Sub-Saharan Africa and Latin America and the Caribbean have higher levels of corruption within their politics and economy, which sometimes affects their overall happiness score report. In general, there is a negative relationship between perceptions of corruption and scores, meaning that the less corruption perceived within a country, the more common a higher happiness score is. However, there are some instances of groups of outliers such as in Sub-Saharan Africa. There is a clustering of countries which have very low levels of corruption, yet they still have low happiness scores.



Generosity vs. Happiness

Finally, we can see that there isn't a very clear relationship between a country's generosity level and happiness score. Southern Asian countries do tend to have a higher level of generosity, which could potentially be due to their cultural customs.



Modeling

Defining RMSE

Now that we have a better understanding of the dataset, we can begin to build models of varying complexity and measure their success using RMSE. This section details the development of 4 different machine learning models, using various teachniques and predictors.

First, we develop a baseline model using just the score averages in order to have a starting point to compare our more advanced models against. We will assess this model (and all future models) using the RMSE (root mean square error), which quantifies how far the predicted values are from the observed values in the final hold out validation set. Therefore want to minimize the RMSE value. The RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2}$$

with y_i as the happiness score for Country i and $\hat{y}_i$ to denote our prediction.

Train & Test Sets

This step creates a data partition on our world happiness data, splitting the data with 75% as a training set and 25% reserved as a final testing set. Since our dataset has 1704 observations, when we split the data both the training and test sets will have a relatively large amount of observations, resulting in less bias and variance. This validation approach will be used for all of our models except the first linear model, which uses 10-fold cross-validation to improve the accuracy slightly. Then, a RMSE function is created based on the formula listed above.

```
test_index <- createDataPartition(happiness_data$Score, times = 1, p = 0.25, list = FALSE)
test_set <- happiness_data[test_index, ]
train_set <- happiness_data[-test_index, ]

#Replace test set NAs with column means from our training set
col_means <- lapply(train_set[,2:11], mean, na.rm = TRUE)
test_set <- replace_na(test_set, col_means)
```

```
#Create Root Mean Squared Error (RMSE) function
rmse <- function(true_scores, predicted_scores){
  sqrt(mean((true_scores - predicted_scores)^2))}
```

Baseline Model

This baseline model simply predicts a given country's score as the average score of the entire dataset. Therefore, regardless of the country or other variable values, each happiness score prediction is the same. This baseline model using the mean happiness score of the dataset is slightly better than randomly guessing, however we know from our data exploration that we can acheive a higher accuracy.

```
#Calculate the average happiness score of all countries in the training set
avg_score <- mean(train_set$Score)

#Calculate RMSE using the average score as the predicted value for each country
#and the scores from the final hold out validation set as the observed values
rmse(test_set$Score,avg_score)
```

```
## [1] 1.096909
```

Model 1: Linear Model

From our exploratory analysis of the dataset, we know that there were very strong relationships with several predictor variables and a country's happiness score. Therefore we can build a linear model to reflect these relationships.

Since our dataset only has 1704 observations, the standard 75%/25% split into training and testing data could be improved upon for the linear approach. This is because our test set will be relatively small and thus our measure of model performance may be weakened. Instead, we can use the method of cross-validation to build K different models that make predictions on all of our data. K-fold cross-validation splits the data into K different parts, with each model trained on K-1 different parts and tested on the remaining part. This process is repeated until each of the K subsets has served as the test set, then the K prediction errors are averaged. This step creates the data partitions on our world happiness data using 10-fold cross-validation. In general, values of K=5 or K=10 have been shown to yield test error estimates that don't have excessively high bias or variance, which is why we have selected K=10.

```
#Define train control for 10 fold cross validation
set.seed(1)
train_control <- trainControl(method="cv", number=10)
```

```
#Train a linear model using 10-fold cross validation
#and make predictions on each of the k subsets
model1 <- train(Score~Log.GDP+Social.support+Healthy.life.expectancy+Freedom+
                Perceptions.of.corruption+Generosity+Year+Region,
                na.action=na.exclude,
                data=happiness_data,
                trControl=train_control, #Generate training and test sets, with cross validation
                method="lm")
#Compute the prediction error RMSE
print(model1)
```

```
## Linear Regression
##
## 1704 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1365, 1364, 1364, 1365, 1365, 1364, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.5267627  0.7863632  0.4054666
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

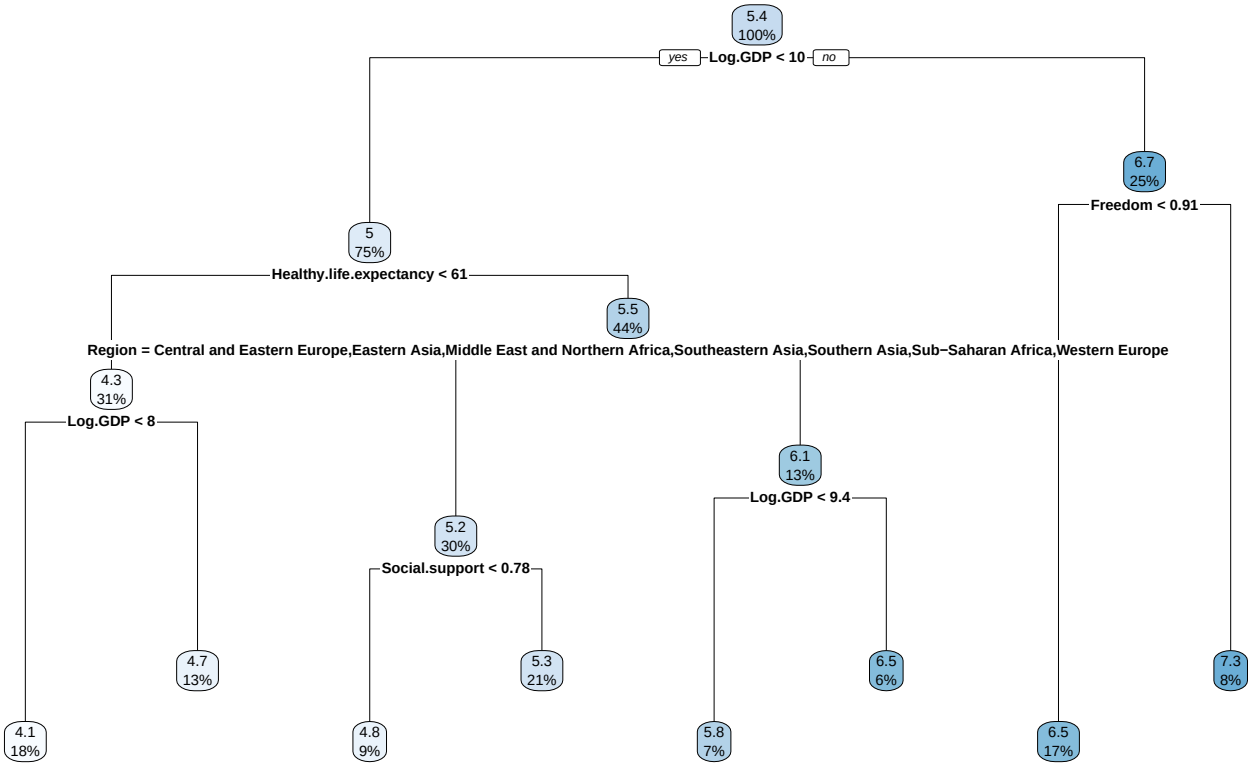
Model 2: Decision (Regression) Tree

Another machine learning technique that can be applied to our dataset is the creation of a decision tree, and in this case a regression tree since the outcome (happiness score) is continuous. To develop this regression tree we will use binary splitting to grow a large tree on the training data that creates partitions of 2 branches multiple times in a top-down algorithm, while simultaneously minimizing the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In this approach, we will apply cost complexity pruning to the tree in order to find the optimal tuning parameter α. The function rpart uses the method of cross-validation and applies a range of cost complexity values to choose this tuning parameter by dividing the training set into k=10 different folds to estimate the test error rate. We can see from the code below that the optimal cost complexity parameter from the cross-validation that was performed was 0.01. The resulting regression tree with all the nodels/splits is also displayed below.

```
#Fully grown regression tree
model2 <- rpart(Score ~ Log.GDP+Social.support+Healthy.life.expectancy+Freedom+
                Perceptions.of.corruption+Generosity+Year+Region,
                method = "anova",
                data = train_set)
#visualize the splits
rpart.plot(model2)
```



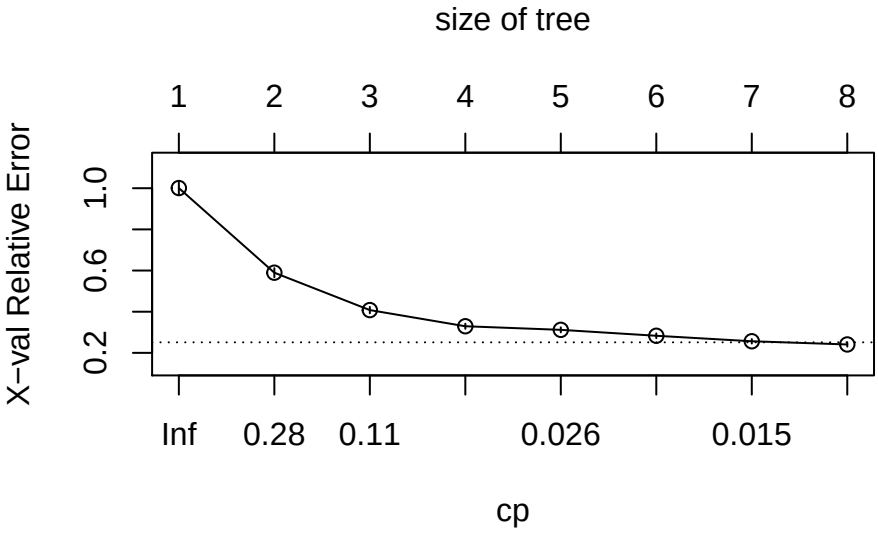

```
#Get score predictions
predicted_scores <- predict(model2, test_set)
#Compute the prediction error RMSE
rmse(test_set$Score,predicted_scores)
```

```
## [1] 0.5591823
```

```
#The most important factors in determining score (according to the features in our model)
model2$variable.importance
```

```
##                   Log.GDP   Healthy.life.expectancy                    Region
##                 975.65188                 811.64198                 782.35050
##            Social.support                   Freedom Perceptions.of.corruption
##                 416.88858                 216.92563                 162.43977
##                Generosity                      Year
##                  29.54836                   2.24254
```

```
#Find the optimal cost complexity parameter that was used in the decision tree
plotcp(model2)
```



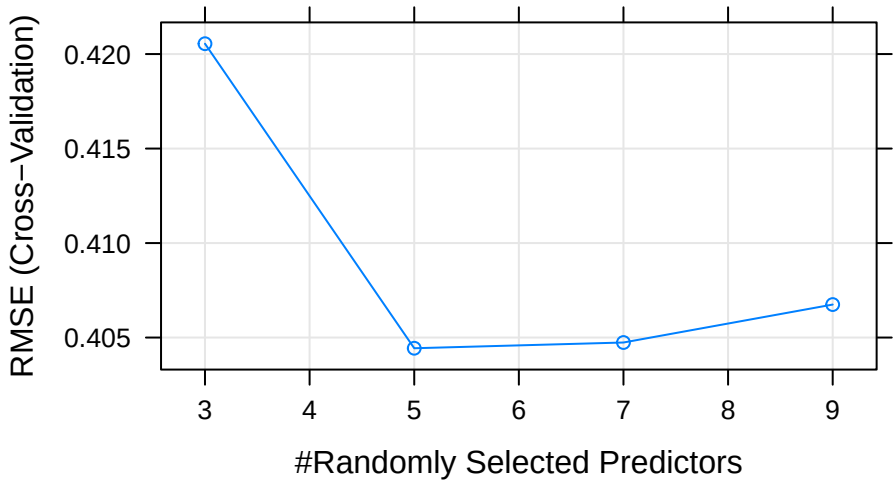
Model 3: Random Forest

Our regression tree model performed worse than our multiple linear regression model because single regression trees tend to have higher variance and therefore poor accuracy. This is because a different training data set can significantly alter the terminal nodes and split values, resulting in completely different predictions. We can improve this accuracy significantly by using a random forest approach.

Random forests enhance predictive accuracy by averaging the results accross lots of different trees. In the code below, each tree is subsequently grown to the fullest extent and 10-fold cross-validation is being used for re-sampling to ensure that the individual trees aren't the same. This results in a large reduction in RMSE from our first 2 models, indicating happiness scores are being predicted with more accuracy.

```
#Train multiple regression trees and average their errors
#use tuneGrid and cross validation to optimize the tuning parameter
model3 <- train(Score~Log.GDP+Social.support+Healthy.life.expectancy+Freedom+
                Perceptions.of.corruption+Generosity+Year+Region,
                method = "rf",
                tuneGrid = data.frame(mtry = c(3,5,7,9)),
                trControl=train_control, #10-fold cross validation to generate different trees
                importance=TRUE, #allows to inspect variable importance
                data = train_set,
                na.action=na.exclude)
```

```r
#Plot the tuning parameter vs. RMSE it achieves
plot(model3)
```




```r
#Find the opitmal tuning parameter mtry
#mtry is the number of variables randomly sampled as candidates at each split in the tree
model3$bestTune
```

```
##   mtry
## 2    5
```

```r
#Final model developed
model3$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 0.159874
##                     % Var explained: 87.76
```

```r
#Make predictions on the test data
predicted_scores <- predict(model3, test_set)
#Compute the prediction error RMSE
rmse(test_set$Score,predicted_scores)
```

```
## [1] 0.4250722
```

Model 4: K Nearest Neighbors (KNN)

Another machine learning approach that could improve the prediction of a given country's happiness score is the non-parametric k-nearest neighbors method (KNN). KNN doesn't explicitly assume any form for our function like with linear regression, which provides a more flexible alternative that can reflect the true curvature of our function.

KNN is a machine learning algorithm that predicts an outcome based on the similiarity with its neighbors, using distance functions. In our case we can use the KNN algorithm to predict a given country's score based on the average score of a grouping of neighboring countries. In the code below we use 10-fold cross-validation in order to find the optimal number of neighbors to use to predict a country's overall well-being. After training the model we see that k=5 is the optimal number of neighboring observations to use to predict a country's happiness score. This results in a successful RMSE value of 0.4791718, illustrating KNN is more accurate that our first 2 models in predicting happiness, however the random forest still outperforms this technique.

```
#First we must split up our training set to create a separate label training set
#Isolate the dependent variable
train_y <- train_set %>% na.exclude() %>% select(Score)
#Remove the non-numeric factors and NA values
train_x <- train_set %>% select(-Score,-Year,-Country,-Region) %>% na.exclude()

model4 <- train(train_x, train_y$Score,
                method="knn",
                trControl = trainControl("cv", number = 10),
                preProcess = c("center","scale"),
                tuneLength = 10)
#Find the optimal value of the tuning parameter (how many neighbors to use)
model4$bestTune
```

```
##   k
## 1 5
```

```
#Make predictions on the test data
predicted_scores <- predict(model4, test_set)
#Compute the prediction error RMSE
rmse(test_set$Score,predicted_scores)
```

```
## [1] 0.4489438
```

Results

Method	RMSE
Baseline Model	1.1112900
Linear Model	0.5267627
Decision (Regression) Tree	0.5568067
Random Forest	0.4107367
K Nearest Neighbors	0.4791718

Conclusion

This report successfully developed 4 different machine learning algorithms to predict a country's overall happiness based on the attributes from the World Happiness Report. Using the cleaned world happiness data that included GDP, social support levels, healthy life expectancy, generosity levels, freedom levels and perceptions of corruption, we were able to successfully build a random forest algorithm with a low RMSE of 0.4107367.

The random forest model produced significantly better RMSE results than the first 2 models and the k-nearest neighbors model was the only other model that produced a comparable RMSE. The k-nearest neighbors approach produced a RMSE of 0.4791718 and runs faster than the random forest algorithm since calculating distances is much less time consuming than growing lots of full regression trees.

Overall the random forest algorithm, which includes all the features in the world happiness dataset performed the best (resulted in the smallest RMSE) of all the models built.

Limitations

This project was constrained by computer memory and time limitations. There are many more modeling/sampling techniques that could be applied with improved hardware that would lead to even better results.

Future Work

Further exploration of ensemble methods and other machine learning solutions could potentially improve the prediction results. Other potential options for improving the RMSE include, but aren't limited to, expanding the dataset (finding data from before 2005 and/or more predictors), examining/potentially removing outliers and performing more detailed feature selection. All in all, this project provides very strong predictions for overall happiness of countries across the world and contains interesting data that can be further explored and analyzed.