World Happiness Prediction Models

HarvardX (edX) Data Science Professional Certificate: PH125.9x CYO Capstone

Joseph Thomas

May 03, 2021

Contents

1 Introduction

1.1 Background

The *World Happiness Report 2021* (hereafter, WHR) is the ninth such report and is the focus of this assigned project (HarvardX (edX) Data Science Professional Certificate: PH125.9x Choose Your Own Capstone project).

The WHR is a landmark survey of the state of global happiness. The happiness scores and other data derive from the annual Gallup World Poll. The report continues to gain global recognition as governments, organizations, and civil society increasingly use happiness indicators to inform their policy-making and operational decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of overall well-being, also known as happiness, can be used to assess the progress (or decline) of nations. The reports review the state of happiness across the globe and show how the new science of happiness explains personal and national variations in happiness. The full report is available for easy online reading or download (pdf) at: https://happiness-report.s3.amazonaws.com/2021/WHR+21.pdf

The 2021 WHR is of particular importance to me as a Human Resources executive consulting on multiple employee populations (move than 60,000 employees globally) in this time of the COVID-19 pandemic. In particular, I am first interested in the effects of COVID-19 on the structure and quality of people's lives, and secondly to understand and evaluate how governments, organizations, and society in general have dealt with the pandemic. Much of that work is beyond the scope of this assigned report; however, this report will serve as the basis for actual analysis and recommendations for the organizations that have engaged my services.

1.2 Dataset Description

As noted above, the WHR is the dataset selected for this project. WHR is one of many datasets available from of Kaggle for data science learning and information (www.kaggle.com). The WHR dataset contains the happiness score of 149 countries around the world paired with various versions of six additional factors for a total of 20 variables (columns) including the name of the nation and its corresponding global region. The seven numeric factors are scores which reflect measures of social support, life expectancy, GDP (*per capita* or economic production), generosity, perceptions of corruption, freedom in making life choices, and dystopia residual which we'll explain shortly.

The variables that contribute to the happiness score (primarily those labeled, "explained by. . . ") estimate the extent to which each of six factors contribute to making perceptions of happiness higher in each country than they are in Dystopia, a hypothetical country that has values equal to the lowest national averages for each of the six factors. Dystopia_residual, which is something of a seventh variable, values have no impact on the total happiness score reported for each country. They do, however, explain why some nations rank higher than others when listed in rank order. In short, the higher value of each of variable and the resulting happiness score, the *happier* the nation.

1.3 Project Goal and Key Steps

The goal of this assignment is to create a report using R Markdown to analyze an available dataset and present the related findings, along with supporting statistics and figures.

Having selected the WHR 2021 as the available dataset, the purpose of this work is to determine which *happiness* factors are more important (correlated) to living a happier life in the context of the nations and regions where people live. In addition, we will explore several machine learning algorithms (models) that can be used to predict happiness scores and compare those models to determine which algorithm is best (or better) suited for use as a prediction tool from datasets like the WHR.

Key steps will include loading and tidying the dataset, exploring and visualizing the data, developing and testing multiple prediction models, and assessing the results and model performance.

2 Methods and Analysis

2.1 Prepare the Analytical Environment (RStudio)

In order to conduct the necessary analysis and visualizations, a number of R packages will need to be installed and loaded. The packages and libraries included here are the usual pallet or tools I find necessary in my almost daily worklife. There may indeed be other packages that more elegantly accomplish the tasks at hand; nevertheless, these are the packages I most frequently use.

2.2 Load Data

As mentioned above, the *World Happy Report 2021* dataset is available from numerous sources including multiple sources on www.kaggle.com. Many of those files, however, cannot be downloaded without logging into an account or otherwise registering. So for the convenience of those who will read or review this report, I have posted a copy of the three required files (r, rmd, and pdf) to my project Github repository. I recommend that you pull all four files to your RStudio IDE.

My public Github repository can be found at: https://github.com/jrt1403/cyo_capstone.git

The data file is stored as "world-happiness-report-2021.csv"

```
# PLEASE NOTE: This code chunk is set to automatidcally load
# the data file from my public Github repo:
# https://github.com/jrt1403/cyo_capstone.git.  If, however,
# you prever to pull all four files (r, rmd. pdf, and csv)
# from Github directly, I recomment you set this chunk to
# eval = FALSE to prevent the automatic download.

# Original source file: World Happiness Report 2021
```

```
# (www.kaggle.com) Specific file:
# https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021

# Acquire / load data whr21 <-
# read_csv('world-happiness-report-2021.csv')
library(readr)
urlfile = "https://raw.githubusercontent.com/jrt1403/cyo_capstone/main/world-happiness-r
whr21 <- read_csv(url(urlfile))
rm(urlfile)
```

2.3 Review Raw Data Structure

A cursory review of the data set reveals of the dataset contains 149 observations (rows) and 20 variables (columns). The variable names are a bit unwieldy for our purposes or visualizing and tabulating our data. In fact, the whole of the table cannot be legibly displayed or summarized:

Table 1: Impossible to read: Structure of Happiness Dataset

Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
Finland	Western Europe	7.842	0.032	7.904	7.780	10.775	0.954	72.0	0.949	-0.098	0.186	2.43	1.446	1.106	0.741	0.691	0.124	0.481	3.253
Denmark	Western Europe	7.620	0.035	7.687	7.552	10.933	0.954	72.7	0.946	0.030	0.179	2.43	1.502	1.108	0.763	0.686	0.208	0.485	2.868
Switzerland	Western Europe	7.571	0.036	7.643	7.500	11.117	0.942	74.4	0.919	0.025	0.292	2.43	1.566	1.079	0.816	0.653	0.204	0.413	2.839
Iceland	Western Europe	7.554	0.059	7.670	7.438	10.878	0.983	73.0	0.955	0.160	0.673	2.43	1.482	1.172	0.772	0.698	0.293	0.170	2.967
Netherlands	Western Europe	7.464	0.027	7.518	7.410	10.932	0.942	72.4	0.913	0.175	0.338	2.43	1.501	1.079	0.753	0.647	0.302	0.384	2.798
Norway	Western Europe	7.392	0.035	7.462	7.323	11.053	0.954	73.3	0.960	0.093	0.270	2.43	1.543	1.108	0.782	0.703	0.249	0.427	2.580

To make this cursory review easier, here is a look at the column names and count and the observation count:

Number of columns (variables): 149

Number of rows (observations:) 20

Column # and Column Name

```
##  [1] "Country name"
##  [2] "Regional indicator"
##  [3] "Ladder score"
##  [4] "Standard error of ladder score"
##  [5] "upperwhisker"
##  [6] "lowerwhisker"
##  [7] "Logged GDP per capita"
##  [8] "Social support"
##  [9] "Healthy life expectancy"
## [10] "Freedom to make life choices"
## [11] "Generosity"
## [12] "Perceptions of corruption"
## [13] "Ladder score in Dystopia"
## [14] "Explained by: Log GDP per capita"
## [15] "Explained by: Social support"
```

```
## [16] "Explained by: Healthy life expectancy"
## [17] "Explained by: Freedom to make life choices"
## [18] "Explained by: Generosity"
## [19] "Explained by: Perceptions of corruption"
## [20] "Dystopia + residual"
```

Obviously, there are variables used to calculate other variable and many of them are not relevant to our analysis. Also, the variable names are a bit unwieldy for our purposes of visualizing and tabulating our data.This glimpse into the structure of our dataset makes clear that we will need to tidy this dataset before beginning any meaningful analysis.

2.4 Clean the Data

Now that we have loaded and reviewed our dataset's structure, the dataset appears relatively clean; but, we will need to make three significant adjustments to allow the data to work in our visualizing and tabulating efforts. First, we will remove the columns not relevant to our study. Next, we will rename the remaining columns to be more meaningful and a better "fit". Finally, we convert the Region variable from *character* to *factor*.

```
# Delete unnecessary columns
whr21 <- whr21[, -c(4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 20)]

# Change the name of the remaining variables (columns)
colnames(whr21) <- c("Nation", "Region", "Happiness", "Prosperity",
    "Network", "Wellbeing", "Freedom", "Generosity", "Corruption")

# Convert Region to a *factor* from *character*
whr21$Region <- as.factor(whr21$Region)
```

The cleaned and tidied dataset is far easier to read and review. Our tidied dataset means we are now ready to proceed with our analysis. Now we see the final structure of our dataset consists of 149 observations and 9 variables. Region is reframed as a factor variable and the remaining variables are in numeric form. This configuration situates us nicely to proceed with our analysis.

Table 2: Tidied Structure of Happiness Dataset (First Six Rows)

Nation	Region	Happiness	Prosperity	Network	Wellbeing	Freedom	Generosity	Corruption
Finland	Western Europe	7.842	1.446	1.106	0.741	0.691	0.124	0.481
Denmark	Western Europe	7.620	1.502	1.108	0.763	0.686	0.208	0.485
Switzerland	Western Europe	7.571	1.566	1.079	0.816	0.653	0.204	0.413
Iceland	Western Europe	7.554	1.482	1.172	0.772	0.698	0.293	0.170
Netherlands	Western Europe	7.464	1.501	1.079	0.753	0.647	0.302	0.384
Norway	Western Europe	7.392	1.543	1.108	0.782	0.703	0.249	0.427

For reference and clarity, the following table lists the new column names, the original name

(as given in the dataset), and then a description of the information conveyed:

New Column Name	Original Column Name	Description
Nation	Country name	Names of nations reporting the observations.
Region	Regional indicator	Loosely correlates to the geographic groupings (not exactly by continent) or grouping of similarly situated countries where geography doesn't serve.
Happiness	Ladder score	Rating or measure of happiness on a scale of 0 to 10.
	Explained by...	
Prosperity	Log GDP per capita	GDP (gross domestic product) is the dollar value of goods and services produced a nation in a given year. Per capita GDP is a calculated measure of GDP divided by the nation's population (count). Per capita GDP is a common and useful measure used to compare countries as provides a normalized (per individual) standard measure.
Network	Social support	Means having friends and other people, including family, to turn to in times of need or crisis to give you a broader focus and positive self-image. Social support enhances quality of life and provides a buffer against adverse life events.
Wellbeing	Healthy life expectancy	The cumulative impact of personal health and social constructs that determine how long an individual will live.
Freedom	Freedom to make life choices	The degree of freedom the citizens of a country hold in terms of making their own life decisions without government involvement.

New Column Name	Original Column Name	Description
Generosity	Generosity	Generosity is a quality that's a lot like unselfishness. Someone showing generosity is happy to give time, money, food, or kindness to people in need.
Corruption	Perceptions of corruption	The degree to which dishonest or fraudulent conduct by those in power impacts everyday life of a nation's citizens.
Dystopia+	Dystopia + residual	Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors, is combined with Residuals to serve as a leveling factor that is not absolute zero which would be erroneously low. The Happiness Score depicts the difference between the measured Nation and Dystopia.

2.5 Exploratory Data Analysis and Visualization

2.5.1 Orientation to the Dataset

Happiness (whr21$Happiness) of our happiness dataset is the variable that we hope to train our algorithm to predict from individual or combinations of numeric variables (Prosperity, Network, Wellbeing, Freedom, Generosity, Corruption, and Dystopia+). In that sense, Happiness is also our dependent variable. The remaining numeric variables are the inputs that influence Happiness. Thus, the numeric variables are our independent variables. This perspective of the data set indicates that we will be looking for associations (correlations) between Happiness and the numeric variables as well as combinations of numeric variables. As such, we will begin our data exploration by first considering correlation and regression aspects of the variables.

The dataset contains 149 unique values in the Nation variable meaning that each row or observation represents the values for a single nation. As such, Nation will not serve as our primary grouping factor for comparison. Region, however, will serve somewhat better. The 149 Nations have been sorted by the dataset's author into 10 groups roughly based on 1) geography (not exactly continents), 3) similarly situated in terms of their social and geo-political constructs, or 3) a combination of both. Therefore, we will use Region as our primary grouping key where the Nation variable does not serve.

Also note that there is an "adjustment factor" included in the data set known as the Dystopian Ladder Score. This is the score of that fictitious, lowest-scoring Nation. The value the Dystopian Ladder Score is 2.430 and that value is assigned Nation as well as accounted for in each Nation's overall score. We will use this score in the simplest of our models to be evaluated.

Table 4: Variables, Variable Class, and First Six Observations

Nation	Region	Happiness	Prosperity	Network	Wellbeing	Freedom	Generosity	Corruption
character	NA	numeric	numeric	numeric	numeric	numeric	numeric	numeric
Finland	Western Europe	7.842	1.446	1.106	0.741	0.691	0.124	0.481
Denmark	Western Europe	7.62	1.502	1.108	0.763	0.686	0.208	0.485
Switzerland	Western Europe	7.571	1.566	1.079	0.816	0.653	0.204	0.413
Iceland	Western Europe	7.554	1.482	1.172	0.772	0.698	0.293	0.17
Netherlands	Western Europe	7.464	1.501	1.079	0.753	0.647	0.302	0.384
Norway	Western Europe	7.392	1.543	1.108	0.782	0.703	0.249	0.427

2.5.2 Data Visualization

In this section, we begin to further explore the data set by examining it from various visual perspectives including tabular and graphical representations.

2.5.2.1 Broadest Overview: Distributions of Happiness Perhaps the broadest visual representation of the data set is to examine the number of ratings that fall into each

whole number score category (i.e., 3, 4, 5, etc.).

Figure 1: Distribution of Happiness Score (whole point)

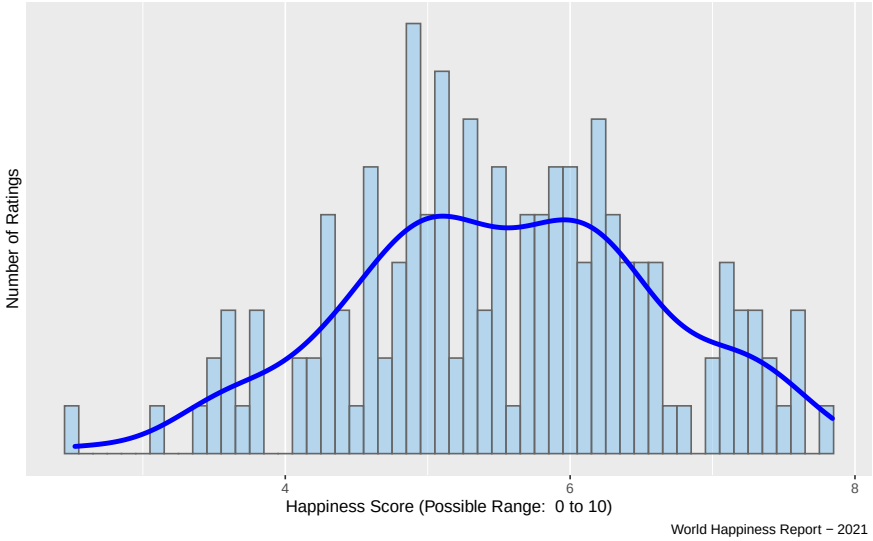


Figure 2: Distribution of Happiness Score (tenth point)

As the Figure 1 and 2 distribution graphics demonstrate, our data ***roughly*** follow a Normal or Gaussian distribution (refer to the dark blue plot lines in the figures above. This will be an important point later as we various prediction models that rely on these distributions: Generalized Linear Model (glm) and Multiple Linear Model (lm) which can be considered a special case of the glm.

Briefly, the difference between a generalized linear model and the linear model in R is about the data's distribution. We assume when we use a lm() in R that our data follow a specific

distribution: Normal or Gauss distribution. Other hand, when using glm() in R, we can specify the data's distribution with the parameters, in most of cases Binomial. glm is an easy way to achieve a linear model when your data don't necessarily follow a Gaussian distribution.

2.5.2.2 Broad Overview of Correlations Between Variables We can understand the relationship among and between the various numeric variables by examining their statistical correlations.

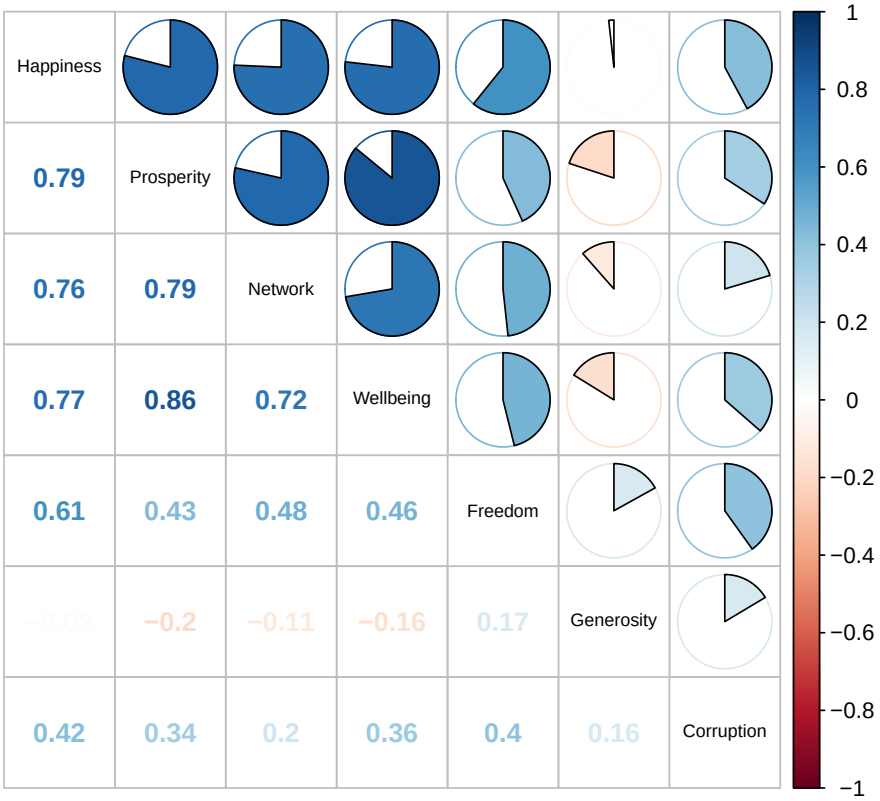


Figure 3: Correlation

Figure 3 demonstrates that correlations between the numeric variables of our data set. The figure indicates that all of the variables, except Generosity, are highly correlated ranging from 0.4 to 0.86 correlation coefficients. Generosity, however, is mildly correlated to the other variables and generally in the negative direction. Generosity may be a useful variable to remove when attempting to tune or improve our prediction algorithm.

Region	Mean	Median	Maximum	Minimum
Central and Eastern Europe	5.984765	6.0780	6.965	5.101
Commonwealth of Independent States	5.467000	5.4715	6.179	4.875
East Asia	5.810333	5.7610	6.584	5.339
Latin America and Caribbean	5.908050	5.9920	7.069	3.615
Middle East and North Africa	5.219765	4.8870	7.157	3.658
North America and ANZ	7.128500	7.1430	7.277	6.951
South Asia	4.441857	4.9340	5.269	2.523
Southeast Asia	5.407556	5.3840	6.377	4.426
Sub−Saharan Africa	4.494472	4.6160	6.049	3.145
Western Europe	6.914905	7.0850	7.842	5.536

Figure 4: Summary: Descriptive Statistics of Happiness Score by Region

2.5.2.3 Overview of Happiness by Region Figure 4 provides a tabular view of the mean and median Happiness scores by Region. Noticeably, means are reasonably similar to means indicating that extreme values are limited. However, there is quite a range between the highs and lows of the maximum and minimum value indicating that some Regions could have a significant spread in their respective ranges.

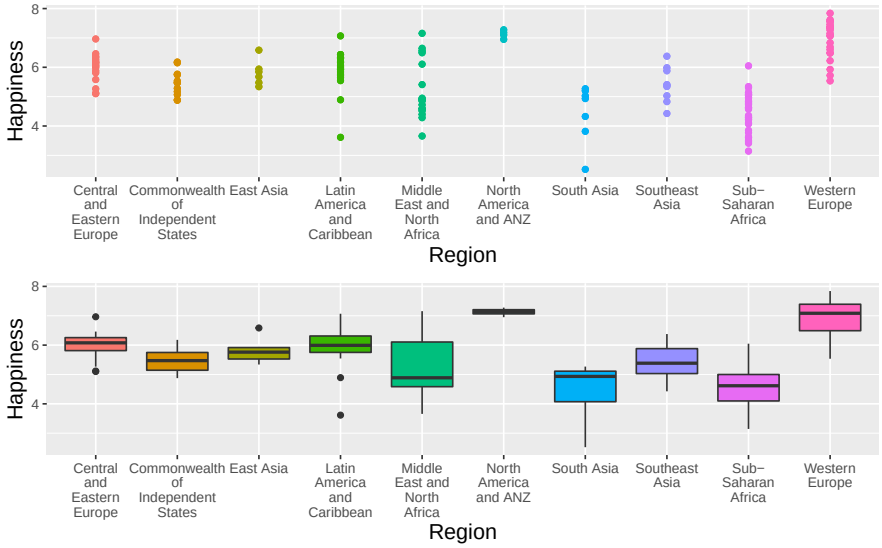


Figure 5: Visual Summary: Happiness Score by Region (Boxplot and Scatterplot)

Figure 5 contains a Box and Scatter plot combined to visually demonstrate the range spread between min and max (scatter plot) as well as the five-number summary of the box plot (i.e., minimum, first quartile, median, third quartile, and maximum). Notably, several regions have a wide range (e.g., Middle East and North Africa, South Asia, Western Europe, and Sub-Saharan Africa) while North America and ANZ (Australia and New Zeland) have a very compact range.

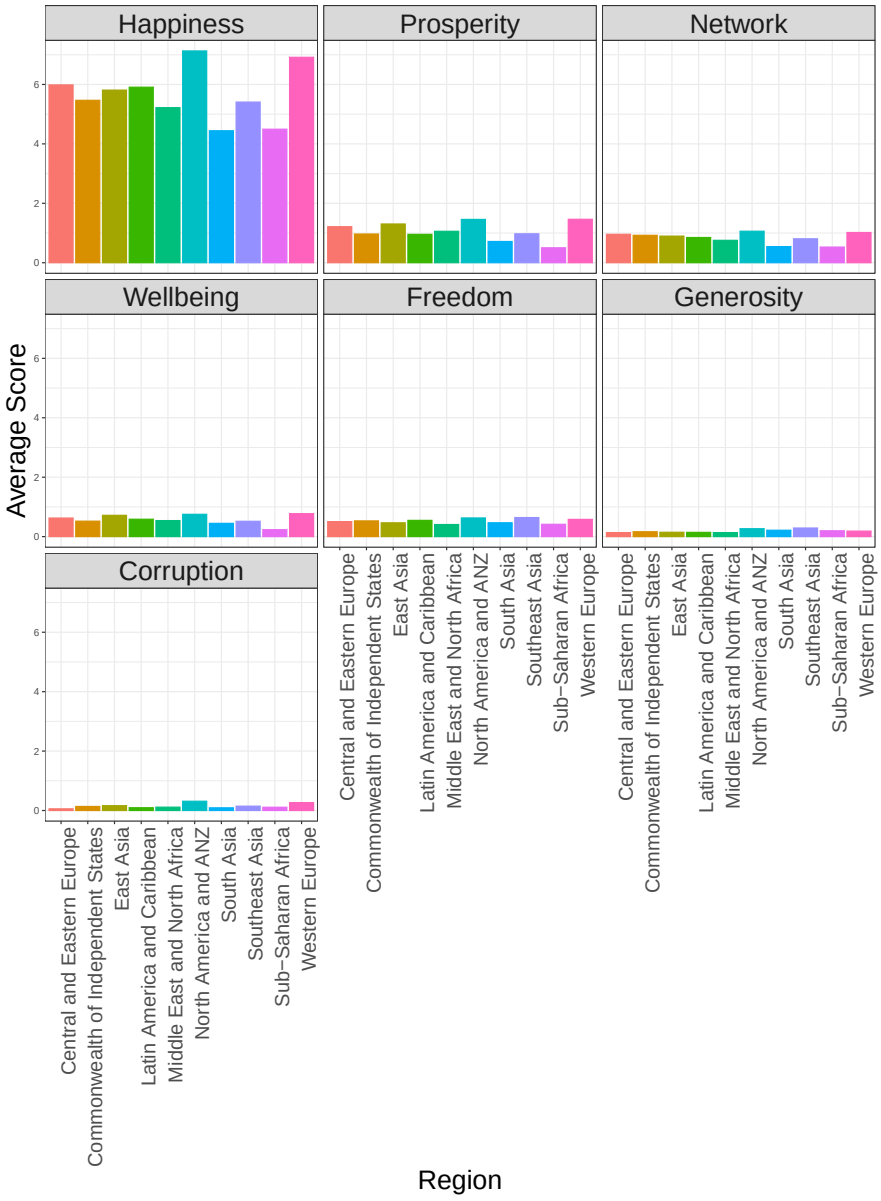


Figure 6: Comparison of All Variables by Region

2.5.2.4 Composite Comparison of All Variables by Region Figure 6 depicts each Region's score by each variable in the WHR. As expected given the highly correlated values, we see spikes for North America and ANZ and Western Europe in each variable and corresponding lows for Southeast Asia and Sub-Saharan Africa. These patterns hold, relatively, for each Region and each variable pair.

In summary, the WHR provides a data set with a dependent variable (Happiness) that roughly follows a normal distribution. The independent variables are highly correlated with one exception (Generosity). There is a wide range of variability between Regions and within each variable's values. As such, this data set lends itself to the development of algorithms to predict the values of Nations absent from the data set of future Nations as this world evolves.

2.6 Development Methods for Multiple Models of Prediction

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from (or near to) the regression line the individual data points fall. In other words, RMSE is a measure of how spread out these residuals are. RMSE can tell us how concentrated the data are around the line of best fit. For this reason, RMSE is commonly used in climatology, forecasting, and regression analysis to verify experimental results. We will use RMSE as our methodology to train and test our models for accuracy in predictions.

2.6.1 Maching Learning Model Development

With our training and testing subsets prepared and set aside, we will begin to develop nine (9) machine learning algorithms and determine the effectiveness of each in predicting future Happiness scores. For the purpose of having a reference point, we have set a Project Target RMSE of 1.0000 meaning that our goal is to produce a ML algorithm that predicts RMSE of 1.0000 or less or within 1.0000 points of the actual Happiness score.

Note, the first two models are simple, calculated models based on the whole of the dataset and do not rely on training and testing subsets. As such, we will address the appropriate split ratio prior to the development of the linear models and others.

2.6.1.1 Simple Average Method The simplest method for predicting ratings based on a known set of values from a dataset is to assume that every observation (Nation in this case) has the same Happiness Score. The actual rating for movie i by user u, $Y_{u,i}$, is the sum of this assumed rating, μ, plus $\epsilon_{u,i}$, representing the independent errors for that distribution.

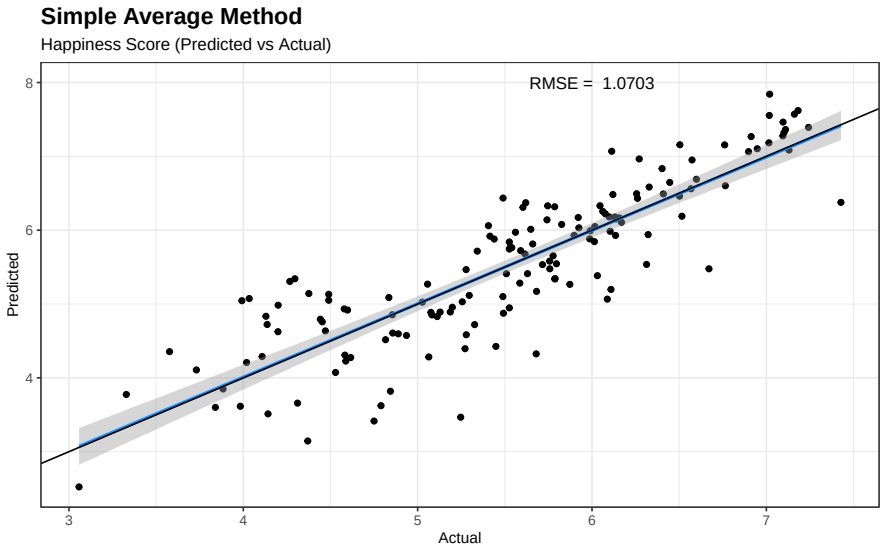
```
# Determine overall average rating for all Nations in train
# set
mu_hat <- mean(whr21$Happiness)
```

We average all Happiness scores to arrive at 5.5328389 as the Happiness score to be assigned to each Nation.

```
# Calculate RMSE between each test set rating included and
# the overall average (mu_hat)
RMSE_simple_average <- RMSE(whr21$Happiness, mu_hat)
MSE_simple_average <- RMSE_simple_average^2
```

Furthermore, we calculate the RMSE for the Simple Average Method to be 1.07031 and the MSE to be 1.14557.

Having calculate the RMSE and knowing the Happiness for each Nation, we can now the predicted Happiness scores and graphically represent those against actuals.



Being the "simplest" approach to creating a predictive model, the Simple Average Method sets the bar or standard against which all subsequent algorithms will be measured - unarguably a low bar. The Simple Sum Methods yields an RMSE of 1.07031 which means that a prediction from this model would be only within approximately 1.0 points from the actual measure which is not a particularly useful estimate and still above our Project Target of 1.0000.

The Simple Average Method plot above demonstrates how widely the points are scattered around the goodness of fit line. It also depicts a perfectly horizontal line reflecting our assumption for this model that all Nations have the same Happiness score.

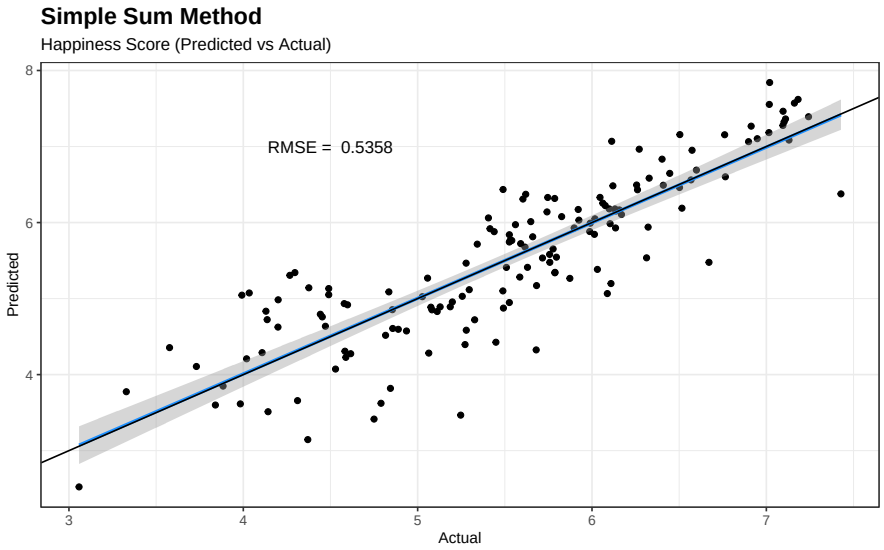
Let's move on to our next "simple" prediction method.

2.6.1.2 Simple Sum Method

The Simple Sum Method is akin to the Simple Average Method. Rather than taking the average of all Happiness scores as the score for each Nation however, the Simple Sum Method assumes that the sum of the independent variables including the Ladder Score for Dystopia equals a Nation's Happiness score. With this assumption in place, we can not predict the Happiness score.

```
# find predicted score by sum method and calculate the
# corresponding RMSE
simple_sum <- whr21 %>%
    mutate(pred_simple_sum = Prosperity + Network + Wellbeing +
        Freedom + Generosity + Corruption + 2.43, RMSE = RMSE(Happiness,
        pred_simple_sum))

# Record RMSE for the simple sum model
RMSE_simple_sum <- RMSE(simple_sum$Happiness, simple_sum$pred_simple_sum)
MSE_simple_sum <- RMSE_simple_sum^2
```



The Simple Sum Method produces am RMSE that is about 50% better than the Simple Average Method. We calculated the RMSEs for the Simple Average Method and the Simple Average Method using calculations based on the whole of the WHR dataset without having to split into training and testing sets. The remaining linear and logistic regression models and and the somewhat more advanced non-linear models will us to split our data set into training and testing sets. What's the best ratio for optimal performance?

2.6.1.3 What is the appropriate dataset split? How then do we then decide what the appropriate proportion or ratio to allocate to each subset? Convention and literature suggest an appropriate split ratio between 70:30 to 90:10 with some caveats regarding very small and very large data sets. We have a very small dataset in the WHR as you may recall with only 149 observations (rows) and 9 variables (columns). We can, however, test for a split that optimizes the RMSE or effectiveness of our algorithm. We accomplish this by calculating RMSE values (our effectiveness and accuracy measure) for a sequence of (relevant but not significant) numbers using a methodology from an earlier assignment. The lowest value of RMSE can then be used to determine the optimal or lowest RMSE prediction.

```
# Test for an 'best' split of full data set (whr21)

# First, we will create sequence of p values or spread to
# test
ps <- (seq(from = 0.2, to = 0.9, by = 0.005))

# Calculate RMSEs for each value of p
rmses <- sapply(ps, function(p) {
    train_index <- as.numeric(createDataPartition(whr21$Happiness,
        times = 1, p = p, list = FALSE))
    train_split <- whr21[train_index, ]
    test_split <- whr21[-train_index, ]
    gof <- glm(Happiness ~ Prosperity + Network + Wellbeing +
```

```
        Freedom + Generosity + Corruption, data = train_split)
    test_split <- test_split %>%
        mutate(pred_score = predict.glm(gof, newdata = test_split))
    RMSE(test_split$Happiness, test_split$pred_score)
})

# Capture the lowest value of RMSE for this split test
low <- min(rmses)

# Capture the split ratio between train and test subsets `r
# x_values[which.min(rmses)]` : `r
# 1-x_values[which.min(rmses)]`
p_train <- ps[which.min(rmses)]
p_test <- 1 - ps[which.min(rmses)]
```

From the basic plot below, the lowest value of RMSE achieved was 0.441364 yielding an optimal data split of 0.74 : 0.26. A few test runs of the following algorithms reveals that the higher we set the training set portion of the ration, the lower our RMSEs. Unfortunately, raising the training portion reduces the test portion and leaves too little data to test. As such and in light of the convention, we will set our split ratio at 70:30 (train_set : test_set).

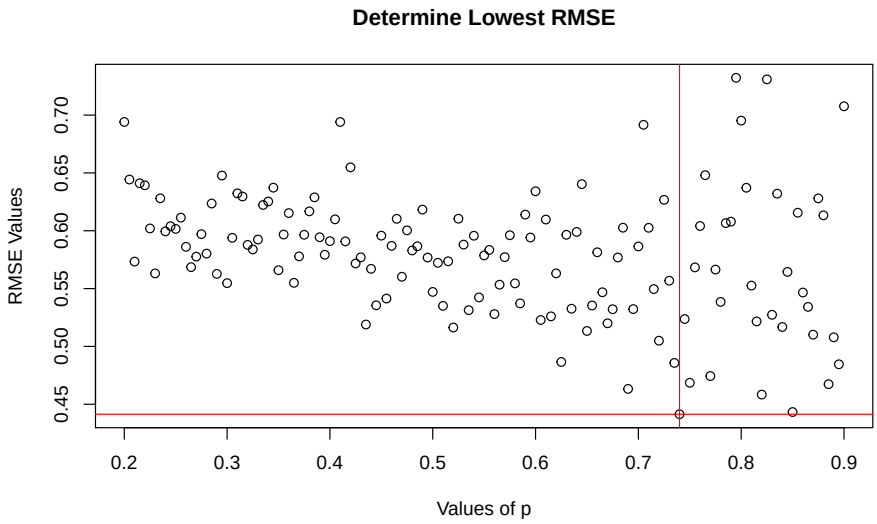


Figure 7: Optimal RMSE

```
## integer(0)
```

2.6.2 Prepare Training and Testing Subsests from WHR at (70:30 ratio)

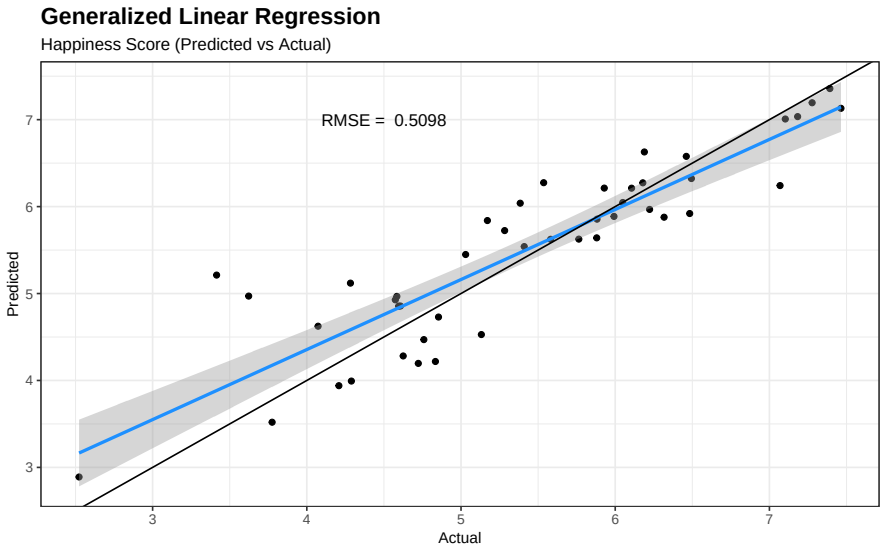
To begin, we will use a slightly less cumbersome method found in the *caTools* package to split the WHR dataset into training and testing subsets. Our variables will remain the same: 1) Dependent variable is the Happiness score; 2) Independent variables are Prosperity, Network,

Wellbeing, Freedom, Generosity, and Corruption; and, 3) Factor / Category variables are Nation and Region

```
# Split the dataset into subsets: train_set and test_set
# using *caTools*
set.seed(1234)
regression_subset <- whr21[3:9]
split = sample.split(regression_subset$Happiness, SplitRatio = 0.7)
train_set = subset(regression_subset, split == TRUE)
test_set = subset(regression_subset, split == FALSE)
```

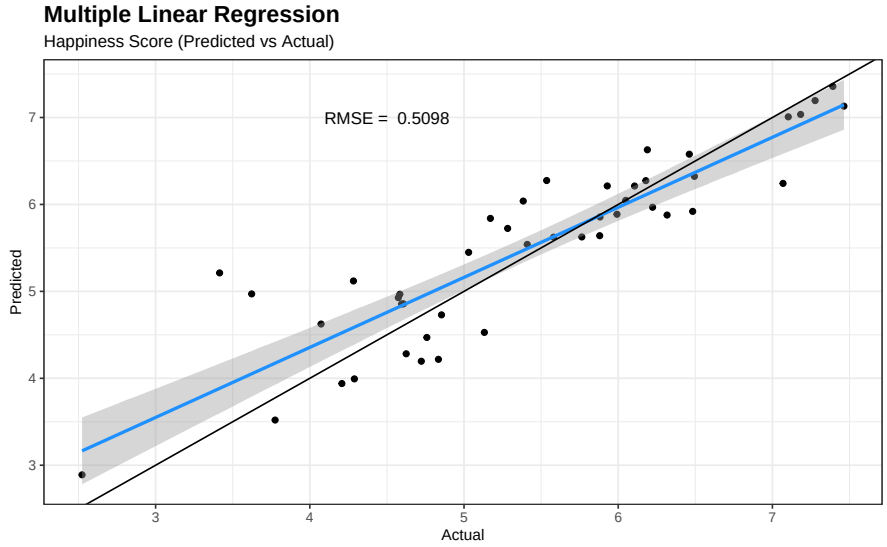
With our dataset split into training and testing subsets, we are not ready to begin training and testing our next seven (7) algorithms. Three of our models are various forms of the generalized linear regression model (generalized linear regression model, multiple liner regression model, and neural net with 0 hidden layers) and the remaining 4 are more advanced models (neural net with 2 hidden layers, support vector regression, decision tree, and random forest model). Let's begin with the glm.

2.6.2.1 Generalized Linear Regression Model (glm) The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable through a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

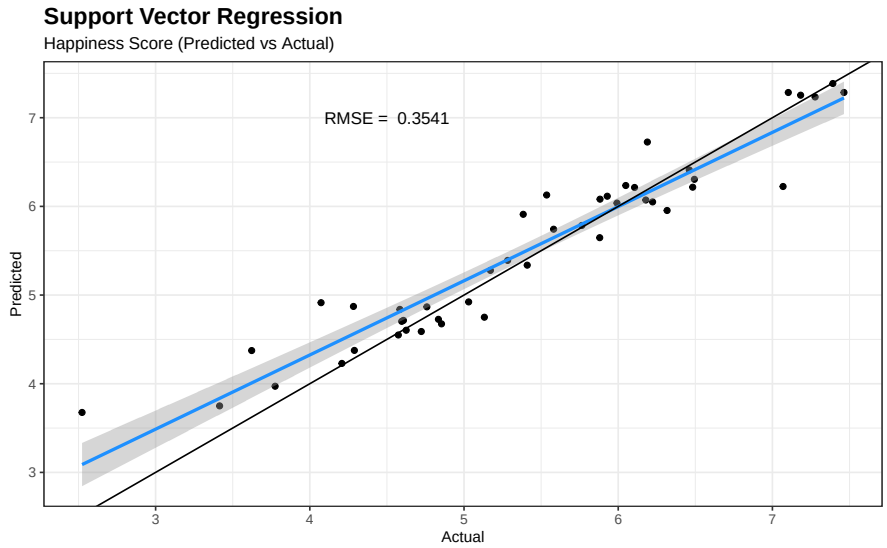


2.6.2.2 Multiple Linear Regression Model (lm) In R as in statistics, using lm() is a special case of glm(). As mentioned above, glm() fits models following the form f(Y) = Xb + e. However, in glm both the function f(Y) (the 'link function') and the distribution of the error term e can be specified. Hence the name - 'generalized linear model'. lm(), on the other hand, fits models following the form Y = Xb + e, where e is Normal (0 , sˆ2).

If the same results are obtained using both lm() and glm() as in our analysis, it is because for glm(), f(Y) defaults to Y, and e defaults to Normal (0, s^2). If the link function and error distribution aren't specified, the parameters that glm() uses produce the same effect as running lm().

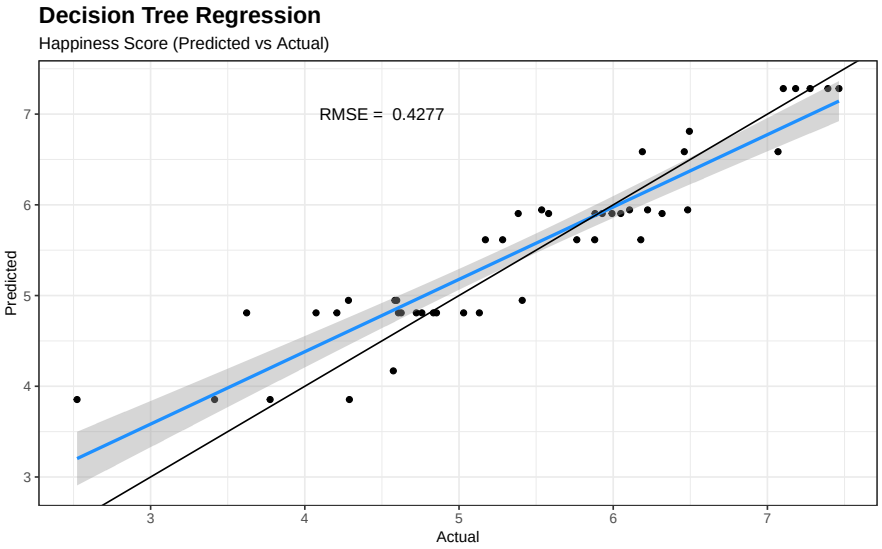


2.6.2.3 Support Vector Regression Model (svm) In machine learning, support vector machines (svm) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. While we use svm here for regression and prediction analysis, they are mostly used in classification problems.

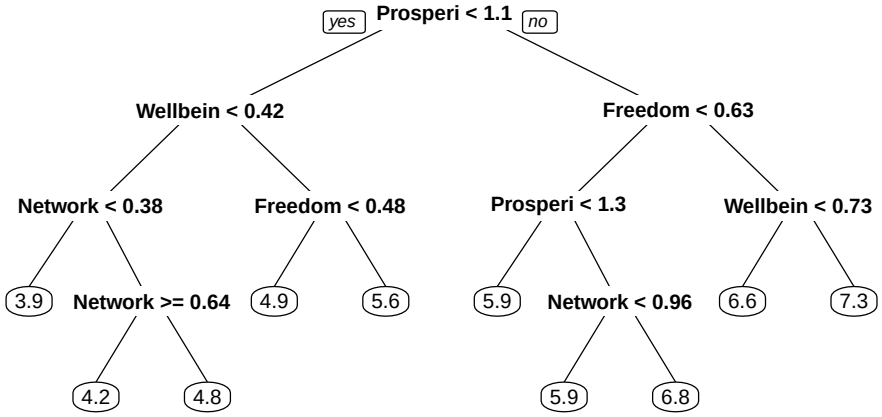


2.6.2.4 Decision Tree Regression Model (rpart) Recursive partitioning is a fundamental tool in data science. It helps us explore the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome or CART for short.. Classification and regression trees can generated through the *rpart package*. Decision Trees are versatile Machine Learning

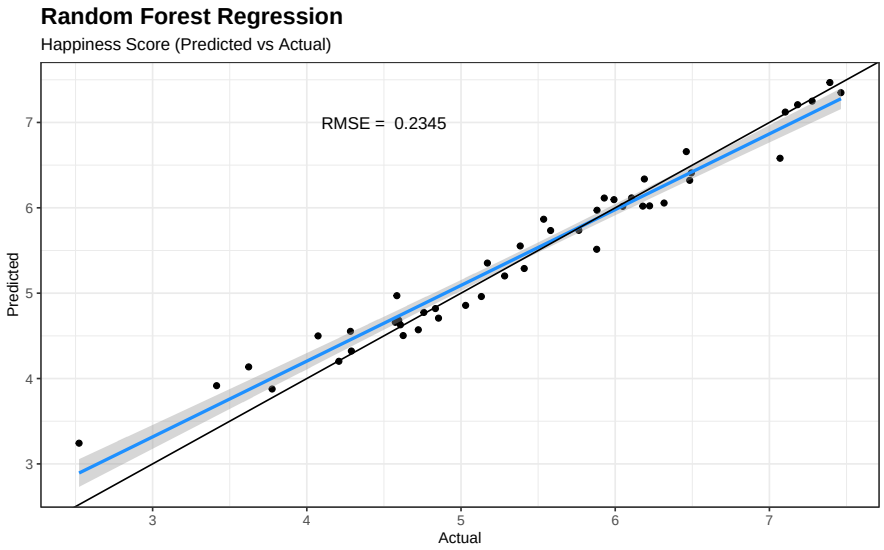
algorithm that can perform both classification and regression tasks. They are very powerful algorithms, capable of fitting complex datasets which is how we use the tool here.



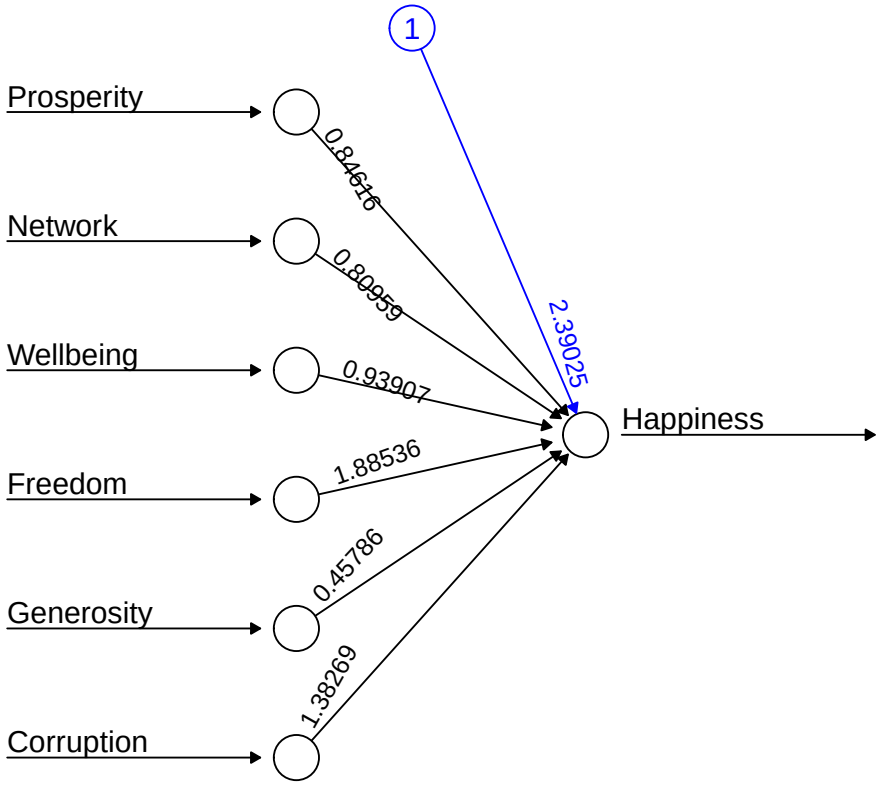
It seems that Decision Tree Regression is not an excellent choice for this dataset. Let's see the tree.



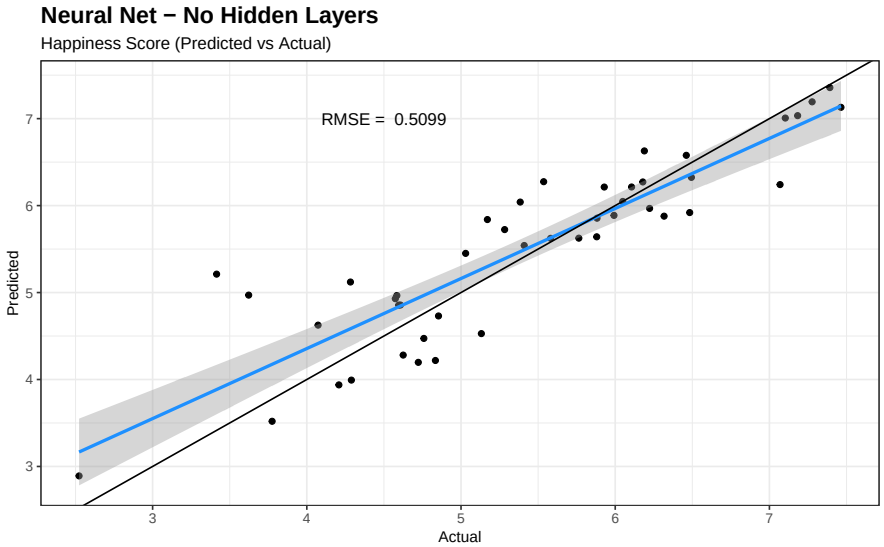
2.6.2.5 Random Forest Regression Model (randomForest) Decision trees are fundamental components of random forests. We use Breiman and Cutler's random forest approach as implemented via the randomForest package in R.



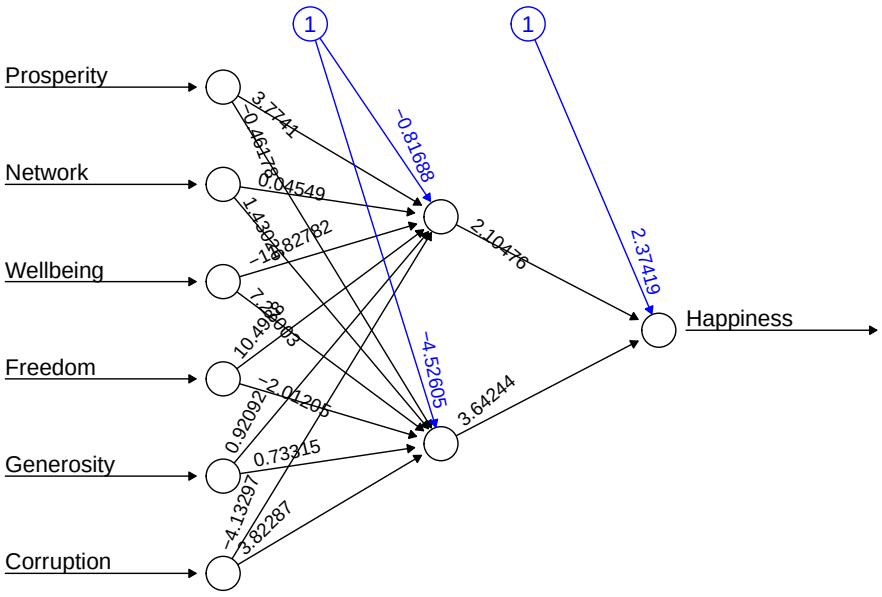
2.6.2.6 Neural Net Model with Zero (0) Hidden Layers (neuralnet) Neural Network (or Artificial Neural Network) has the ability to learn by examples. ANN is an information processing model inspired by the biological neuron system. It is composed of a large number of highly interconnected processing elements known as neurons to solve problems. It follows the non-linear path and processes information in parallel throughout the nodes. A neural network is a complex adaptive system. Adaptive means it has the ability to change its internal structure by adjusting weights of inputs. The number of neurons is defined by the number of hidden layers assigned in R. We provide two example here. One example with 0 hidden layers which causes the results to be very similar to the the linear model results (glm and lm) and another example with 2 hidden layers which produces a similar result as well. Perhaps the linear and normally distributed nature of our data set makes the neural network approach too complex and less effective than alternative algorithms like SVM, Decision Tree, and regression.



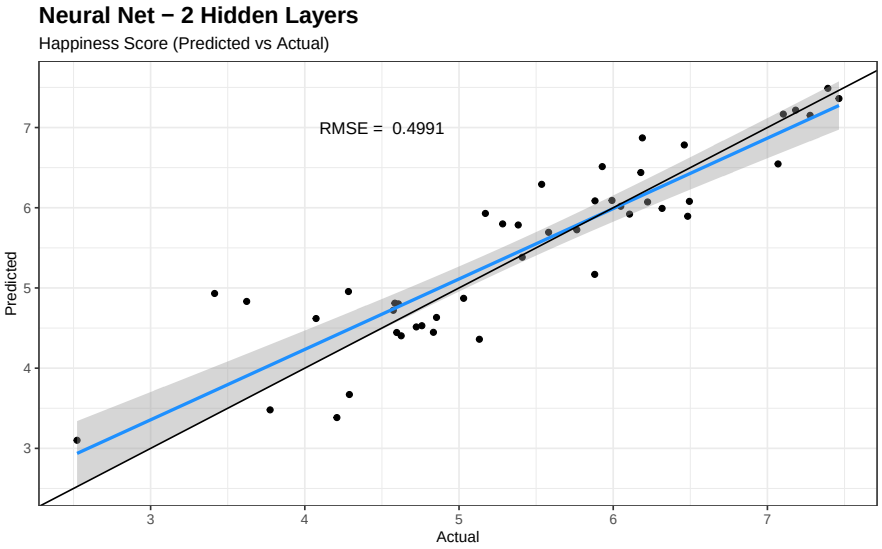
Error: 15.442747 Steps: 588



2.6.2.7 Neural Net Model with Two (2) Hiddeen Layers (neuralnet)



Error: 11.814669 Steps: 5396



Neural networks are more flexible and can be used with both regression and classification problems. Neural networks are good for the nonlinear dataset with a large number of inputs such as images. Neural networks can work with any number of inputs and layers. Neural networks have the numerical strength that can perform jobs in parallel. perhaps the linear and normally distributed nature of our data set makes the neural network approach too complex and less effective than alternative algorithms like SVM, Decision Tree, and regression.

3 Results

We have arranged the results of each method of actual versus predicted values from each of the different machine learning algorithms developed, trained, and tested as part of this exercise. The results are arranged in Figure X below from the least accurate (Simple Average Method) to the most accurate (Random Forest Method).

Actual versus predicted for different machine learning algorithms
Graphic grid view

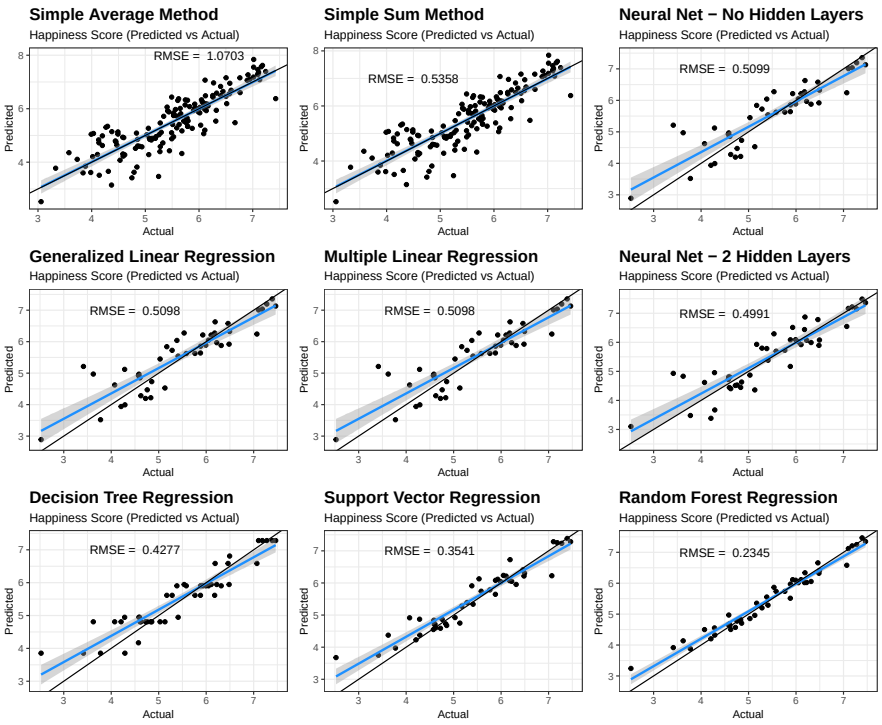


Figure 8: Results Grid

Table 5: Results Table - Target Definition

Method	RMSE	Gap
Project Target	1.00000	-
Simple Average	1.07031	0.07031
Simple Sum	0.53577	-0.46423
Neural Net 0 Hidden Values	0.50988	-0.49012
Generalized Linear Model	0.50976	-0.49024
Multiple Linear Model	0.50976	-0.49024
Neural Net 2 Hidden Values	0.49911	-0.50089
Decision Tree	0.42768	-0.57232
Support Vector Regressison	0.35407	-0.64593
Random Forest	**0.23447**	**-0.76553**

While decision trees are fundamental components of random forests, random forests far outperformed the decision tree methods. Random forests are among the most potent Machine Learning algorithms available today. Random forests improve predictive accuracy by generating a large number of bootstrapped trees (based on random samples of variables), classifying a case using each tree in this new "forest", and deciding a final predicted outcome by combining the results across all of the trees (an average in regression, a majority vote in classification). Generalized Linear Regression, multiple Linear regression, neural net(s), and SVR were distant competitors.

As the results demonstrate, the Random Forest method far outperformed other methods in predicting Happiness scores from the WHR dataset.

4 Conclusion

This report was designed to explore the use of machine learning algorithms and their ability to predict a Nation's Happiness Score after being trained using a relatively small World Happiness Report dataset. In short, we developed, trained, and tested nine (9) machine learning algorithms from the simplest forms of averaging and summing known results to the more complex forms of predicting a categorical (classification tree) or continuous (regression tree) outcome (CART) like Random Forests. The results are remarkable in terms of the dramatic range in the accuracy of the models' prediction (RMSEs).

This report and WHR dataset hold potential impact in the Human Resources arena for predicting "Employee Happiness" as a basis for predicting important workplace factors as retention and turnover, likelihood of injury, employee engagement, and net promoter scores. The two primary limitations for the present study are first the relatively small sample size with fewer than 150 observations and the fact that the underlying dataset is limited to a single year. Future work on this or a similar project could include expanding the Happiness measure to include individuals rather than nations, perhaps a longitudinal approach to including

multiple years of data in a single dataset, and possibly comparing these results to those of more commonly available employee engagement surveys.