The World Happiness Report

Intro to Data Science: Final Class Project

Author: Yiming Huang

UCLA Extension — Fall 2020

Table of Contents

1. Introduction

Perhaps one of the most significant purposes of our life is to be happy. But what actually makes us happy? This question has plagued philosophers for years. Fortunately, researchers have made questionnaires worldwide, called "The World Happiness Report", to further explore the answer to this question through a scientific process. The science of happiness describes personal and national variations in happiness. The report continues to gain global recognition as governments and organizations can use the results to further the policymaking process and assess the progress of countries.

In this project, I used the R language and my main focus is to find out the impact of each factor on happiness. As a result, we can focus on improving these factors to accomplish a higher level of happiness.

1.1 Understanding the dataset

I have chosen the World Happiness Report dataset (2019). In this dataset, we can find happiness rank and score among 156 countries globally based on six factors — GDP per capita, social support, healthy life expectancy, freedom, generosity, and government corruption. The higher the happiness rank, the higher the happiness score. From the shape, we can see that the dataset includes 156 rows and 9 columns. The dataset used for this project is taken from the website Kaggle.com. It is available at the link: https://www.kaggle.com/unsdsn/world-happiness

Detailed description of the dataset using 'summary'

```
# Overall.rank    Country.or.region      Score       GDP.per.capita
# Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
# 1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
# Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
# Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
# 3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
# Max.   :156.00                      Max.   :7.769   Max.   :1.6840
# Social.support  Healthy.life.expectancy Freedom.to.make.life.choices
# Min.   :0.000   Min.   :0.0000          Min.   :0.0000
# 1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
# Median :1.272   Median :0.7890          Median :0.4170
# Mean   :1.209   Mean   :0.7252          Mean   :0.3926
# 3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
# Max.   :1.624   Max.   :1.1410          Max.   :0.6310
# Generosity     Perceptions.of.corruption
# Min.   :0.0000   Min.   :0.0000
# 1st Qu.:0.1087   1st Qu.:0.0470
# Median :0.1775   Median :0.0855
# Mean   :0.1848   Mean   :0.1106
# 3rd Qu.:0.2482   3rd Qu.:0.1412
# Max.   :0.5660   Max.   :0.4530
```

Table 1. Summary of the raw dataset – World Happiness Report 2019

2. Exploratory Data Analysis

2.1 Data Transformation

I observed the structure of the variables using the str() function. The data is quite neat and has no missing values, but I have changed the column names and add new variables "Year" and "Continent" to make it look better and easier for data analysis. The year is 2019. Asia, Africa, North America, South America, Europe, and Australia are the six continents in this dataset. I changed the type of the two variables to factors. The final structure of my dataset has 156 observations and 11 variables. Happiness rank is an integer, Country is a character variable, and the remaining variables are numeric.

2.2 Exploring the data using exploratory graphs

I generated some graphs to explore the relationships among variables in the data provided.

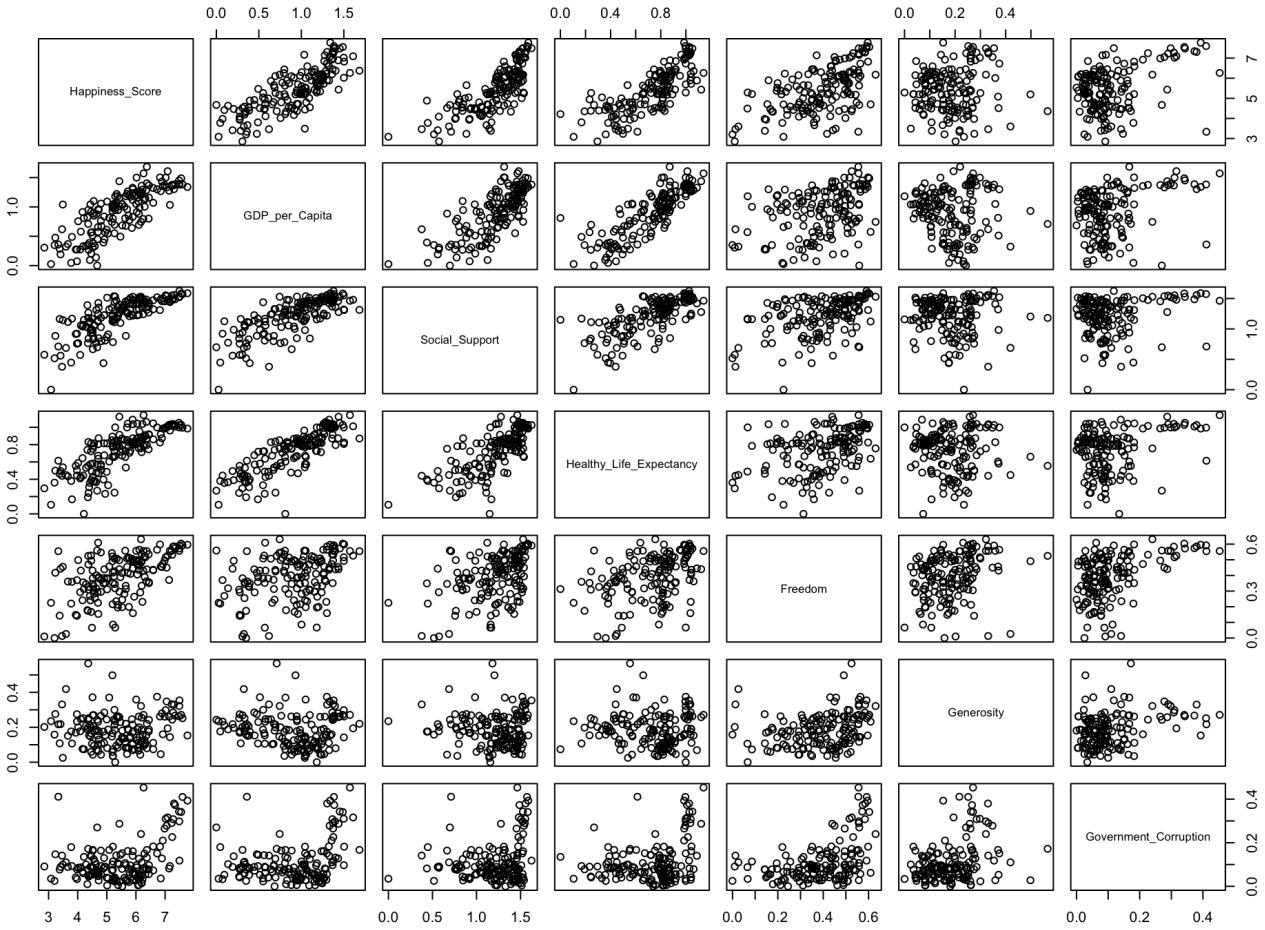
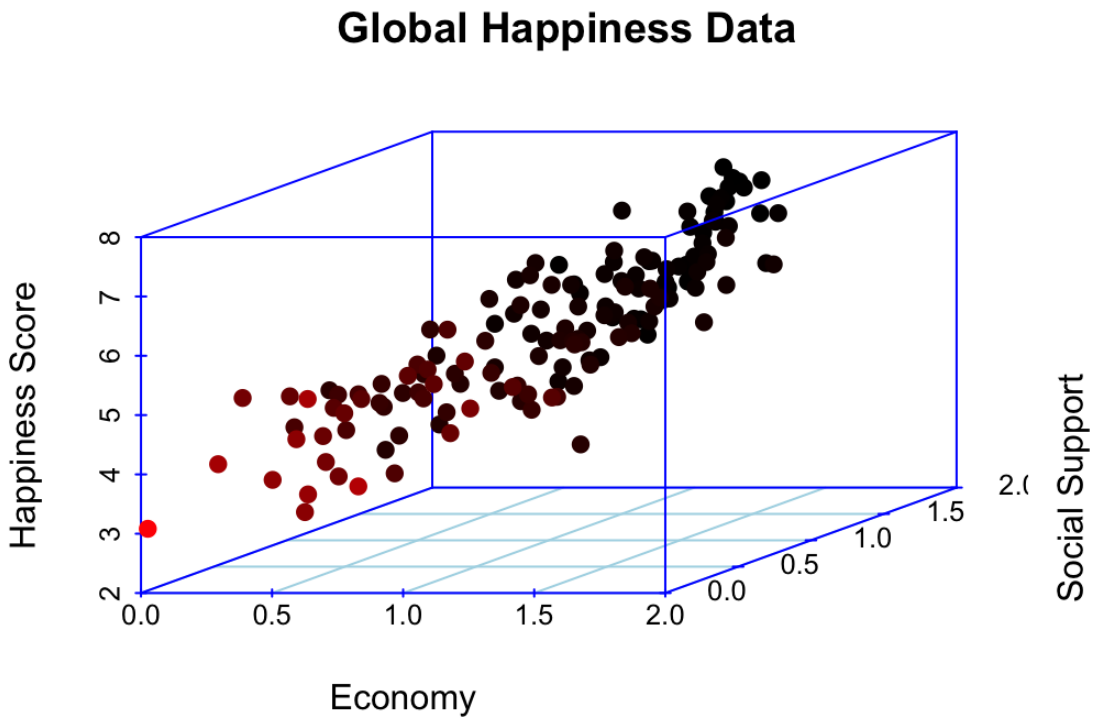
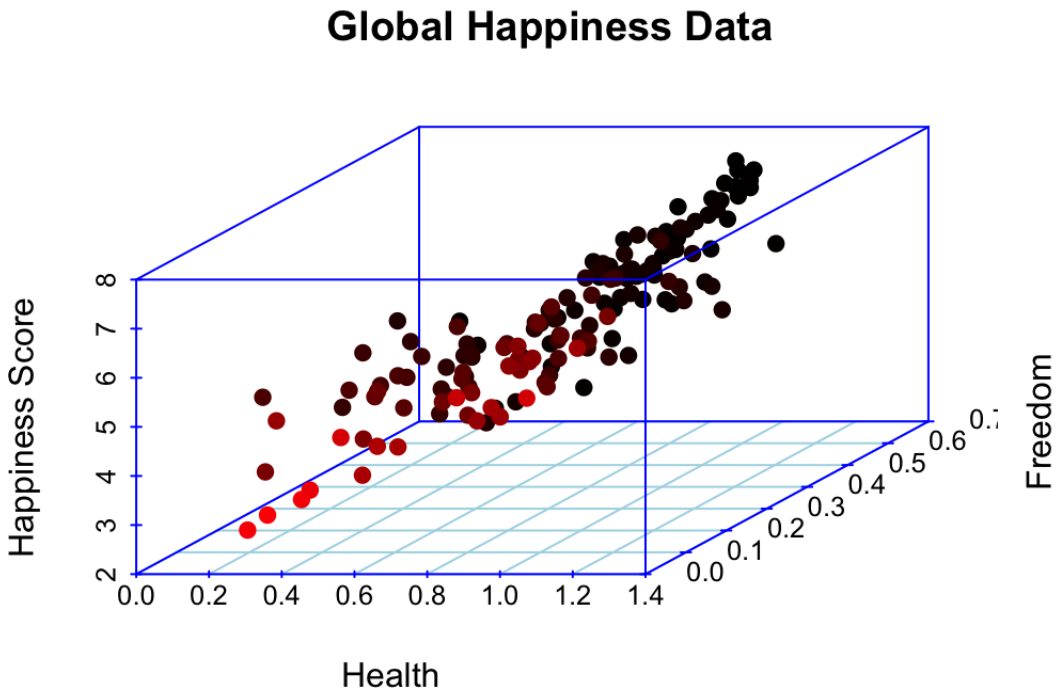


Figure 1. A scatterplot matrix, consisting of scatterplots for each variable-combination

In the scatterplots (Figure 1), I find that GDP, Social Support, Healthy Life Expectancy, and Freedom are highly/moderately correlated with the Happiness score. GDP tends to have the biggest impact on happiness. This is reasonable since adding the variables together provides the Happiness Score.





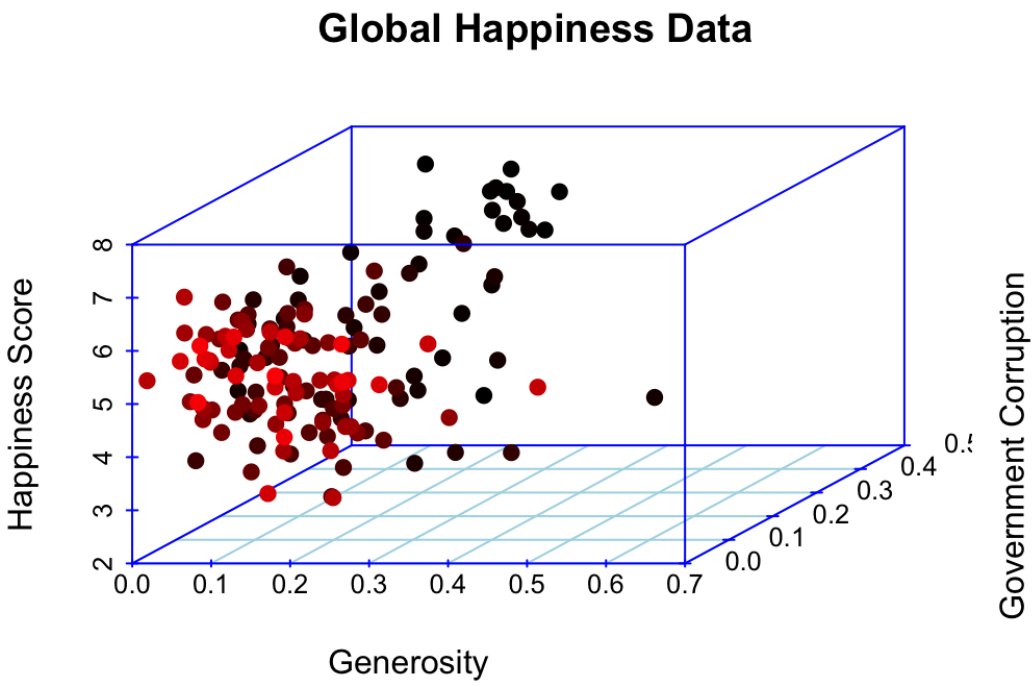


Figure 2. 3-D scatterplots

From the 3D scatterplots (Figure 2), we can see that the better economy and the more social support lead to higher happiness scores. The higher the life expectancy and freedom, the higher the happiness score will be. However, we cannot find linear relationships/trends between generosity/government corruption and higher happiness score.

I also find the correlation of Government Corruption the most interesting here. To further explore this, I conduct Pearson's Chi-squared test and find the two variables are independent because P-value is less than 0.05. When the corruption score is high, however, on the scatterplot it appears to have a positive trend. Thus, Government Corruption seems to be a positive indicator on a certain threshold.

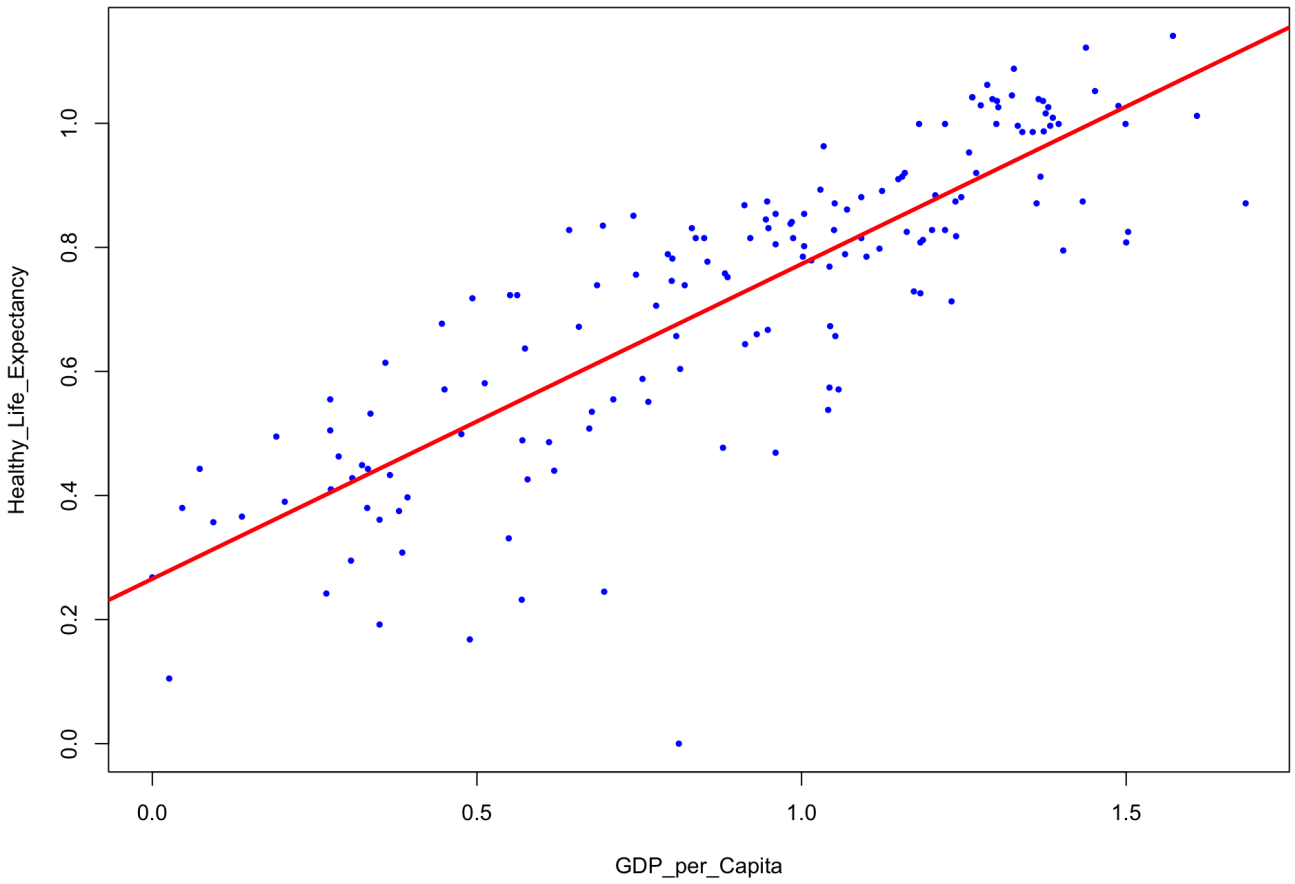


Figure 3. The positive relationship between GDP and Healthy life expectancy

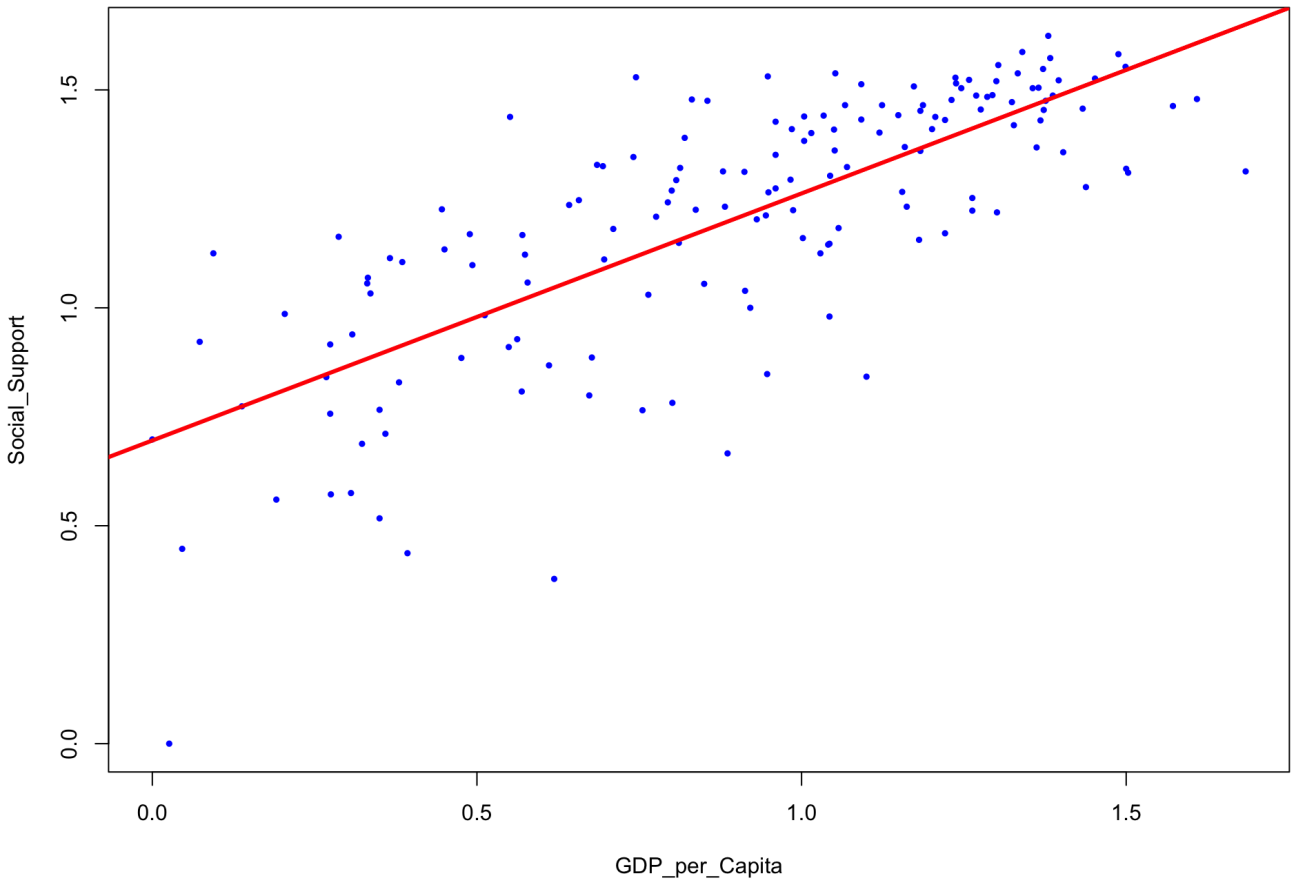


Figure 4. The positive correlation between GDP and Social Support

From Figure 3 and 4, I look at the internal relationships in a little more detail. I chose a couple of variables that interest me: GDP per capita, healthy life expectancy, and social support. As a result, GDP is a significant predictor for Social Support and Healthy Life Expectancy. Probably it is because better economic condition generally results in longer life expectancy as a result of advanced medical services. With a better economic condition/GDP, there will also be a higher rate of social support.

2.3 Exploring different continents regarding individual variable

I further work on the six continents to compare and discover whether there are different trends for them regarding the happiness score.

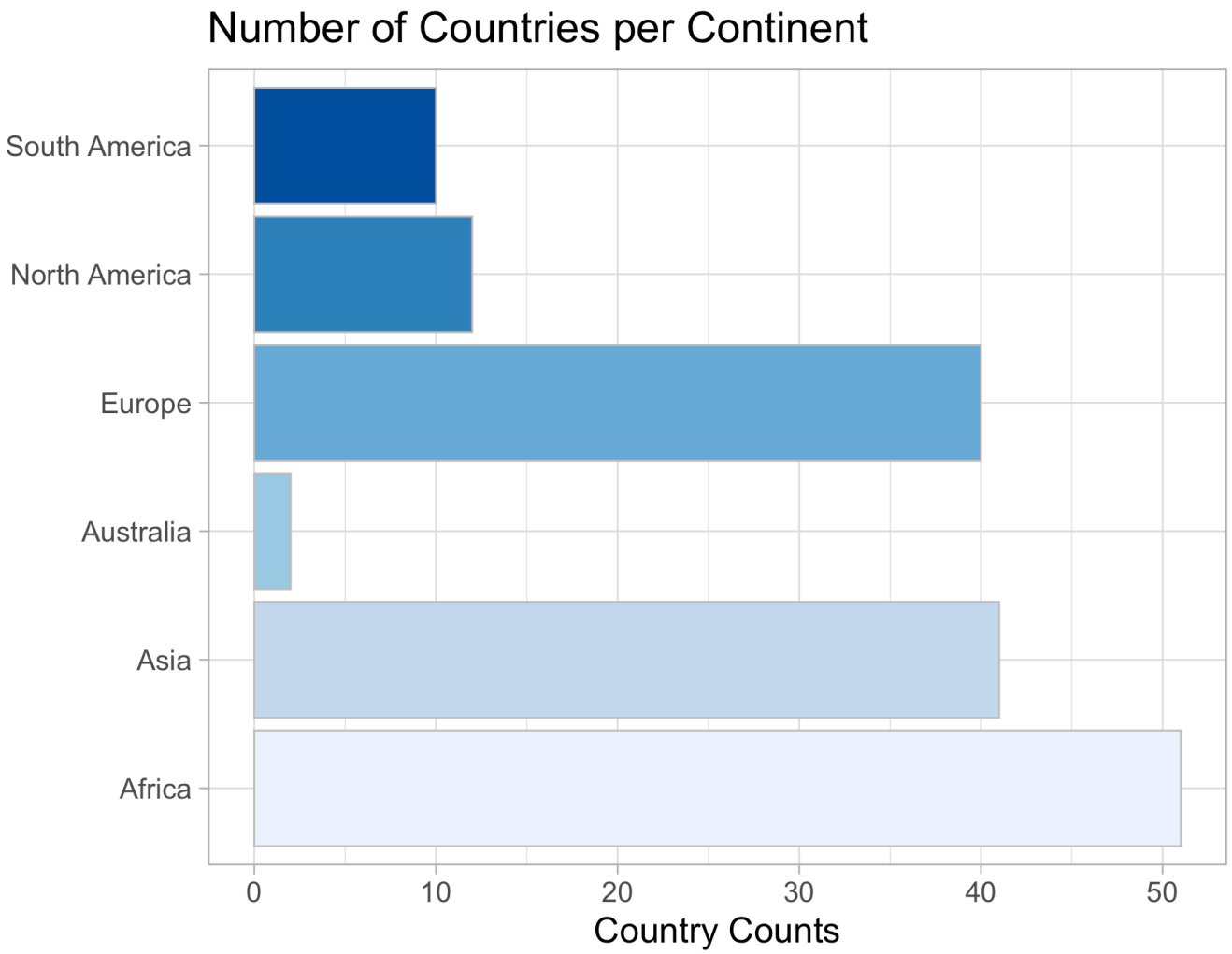


Figure 5. The number of countries per continent

I first count countries grouped in the dataset and compute percentages of observations represented by each continent. Then I visualize it through the bar graph, as shown in Figure 5. There are only two countries in Australia listed in this data file, and around 30 percent of the countries in the dataset are located in Africa.

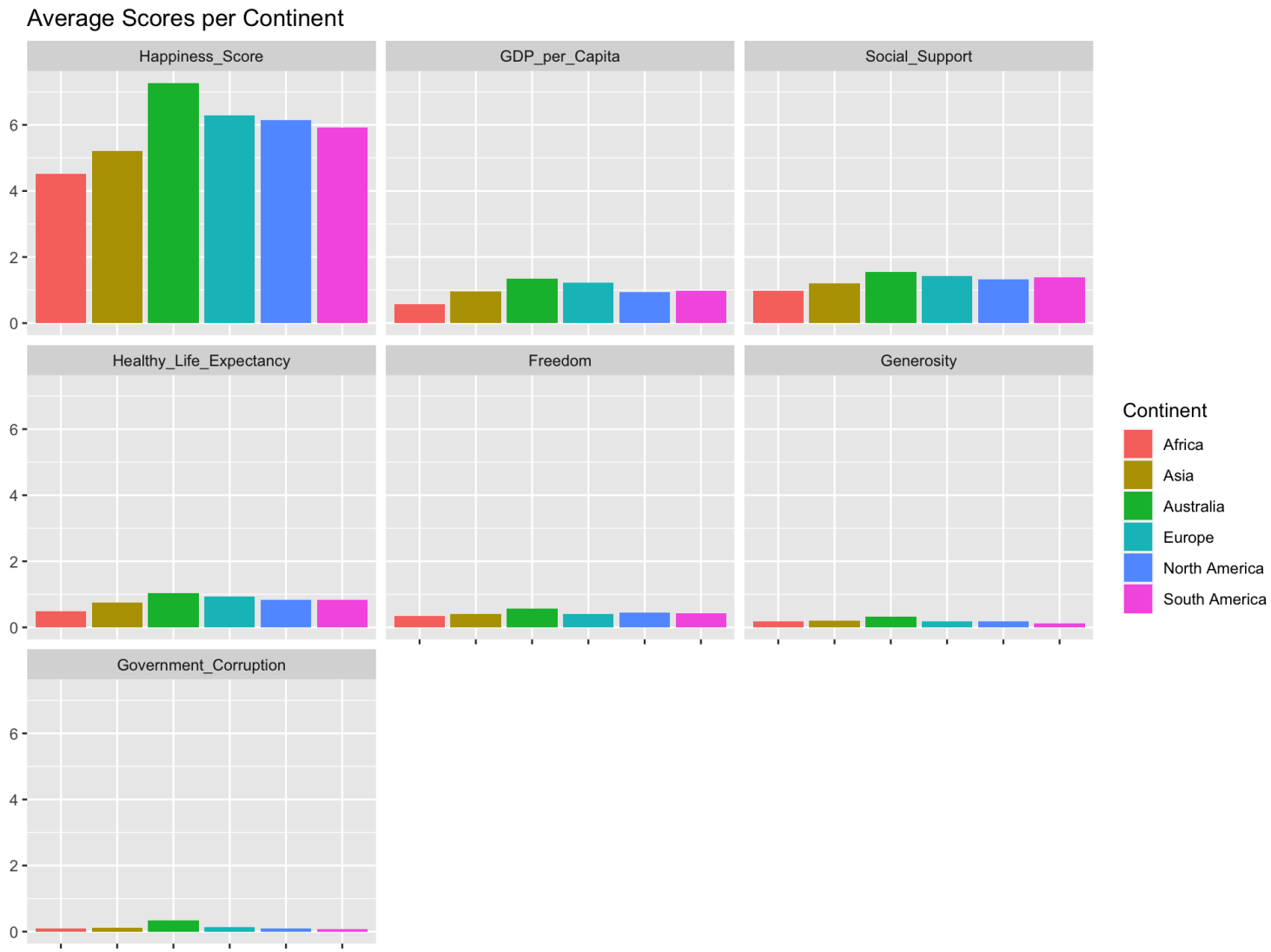


Figure 6. The average score of every variable in different continents

I explore the similarities and disparities regarding each variable for the six continents. From Figure 6, Australia leading in all these categories has the highest average happiness score. The patterns for GDP, social support, and healthy life expectancy are similar. There seem to be few differences among freedom, generosity, and government corruption, with a very limited range of scores. The histograms colored by Continent suggest that generosity and government corruption do not have a big impact on happiness.

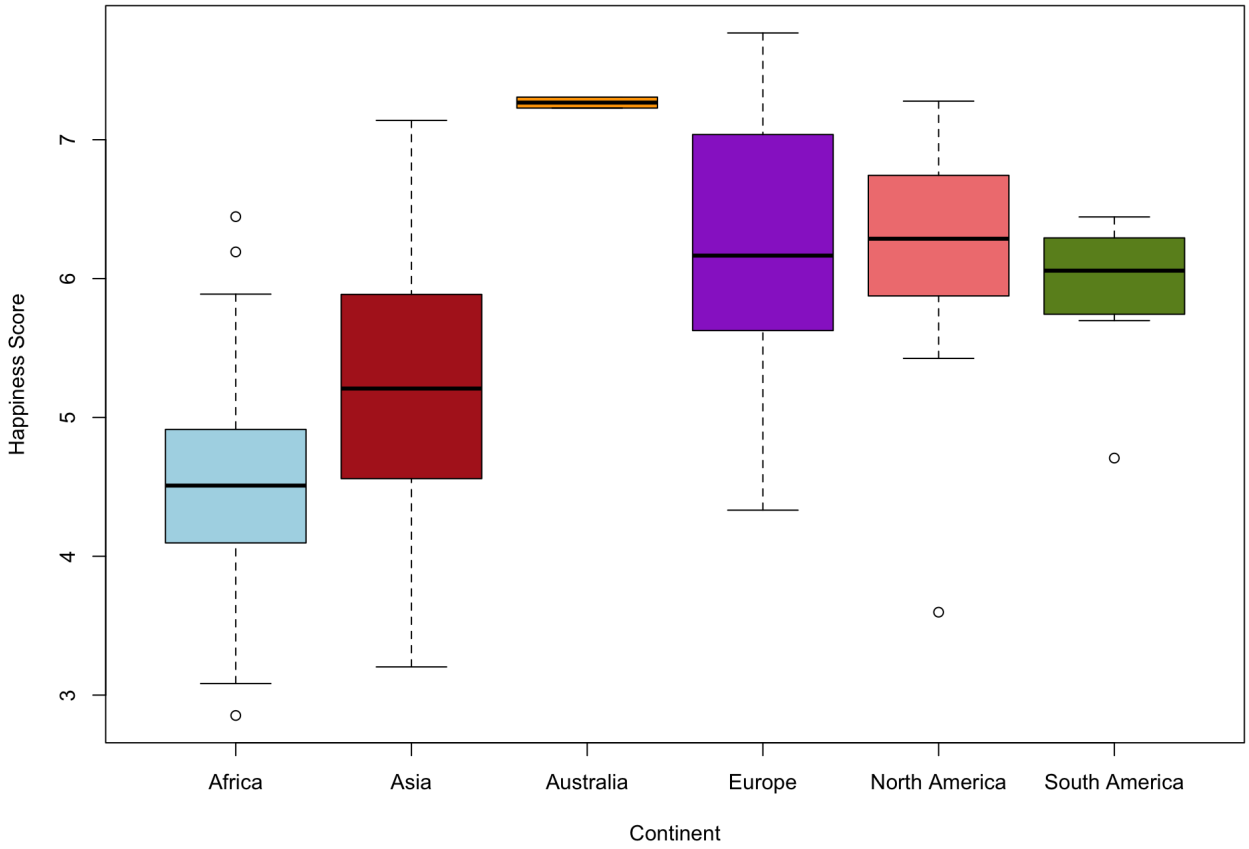


Figure 7. The happiness score distribution in different continents

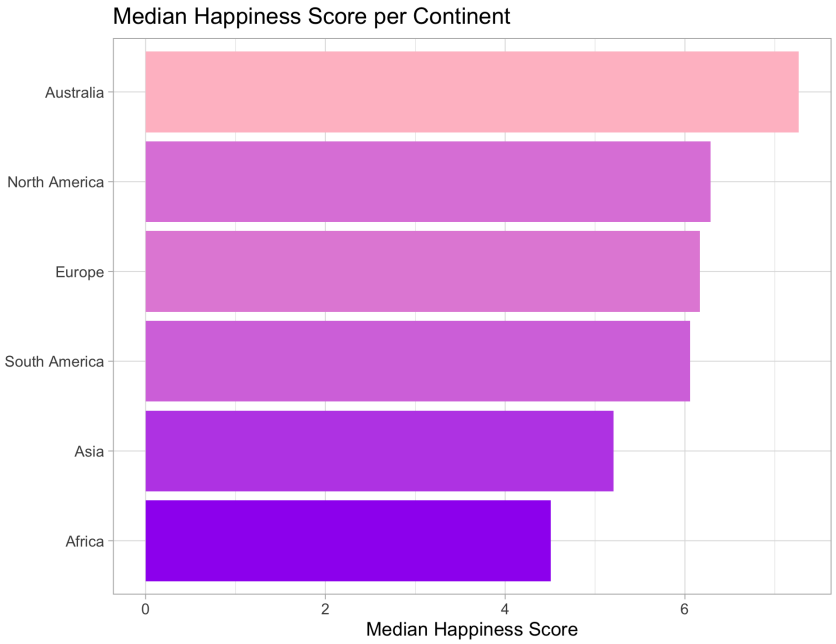


Figure 8. Median happiness score for different continents

As shown in the box plots and bar graph (Figure 7 and Figure 8), Australia has the highest median happiness score, followed by North American and Europe. Africa and Asia have the lowest median happiness score. Besides, Australia has the least range and Asia has the widest range of scores. The country with the highest happiness scores locate in Europe.
Asia and Europe are symmetrical and have no outliers. In Africa, a few outliers are countries with happiness scores higher than 6 and lower than 3. North America and South America have a few outliers with lower happiness scores.

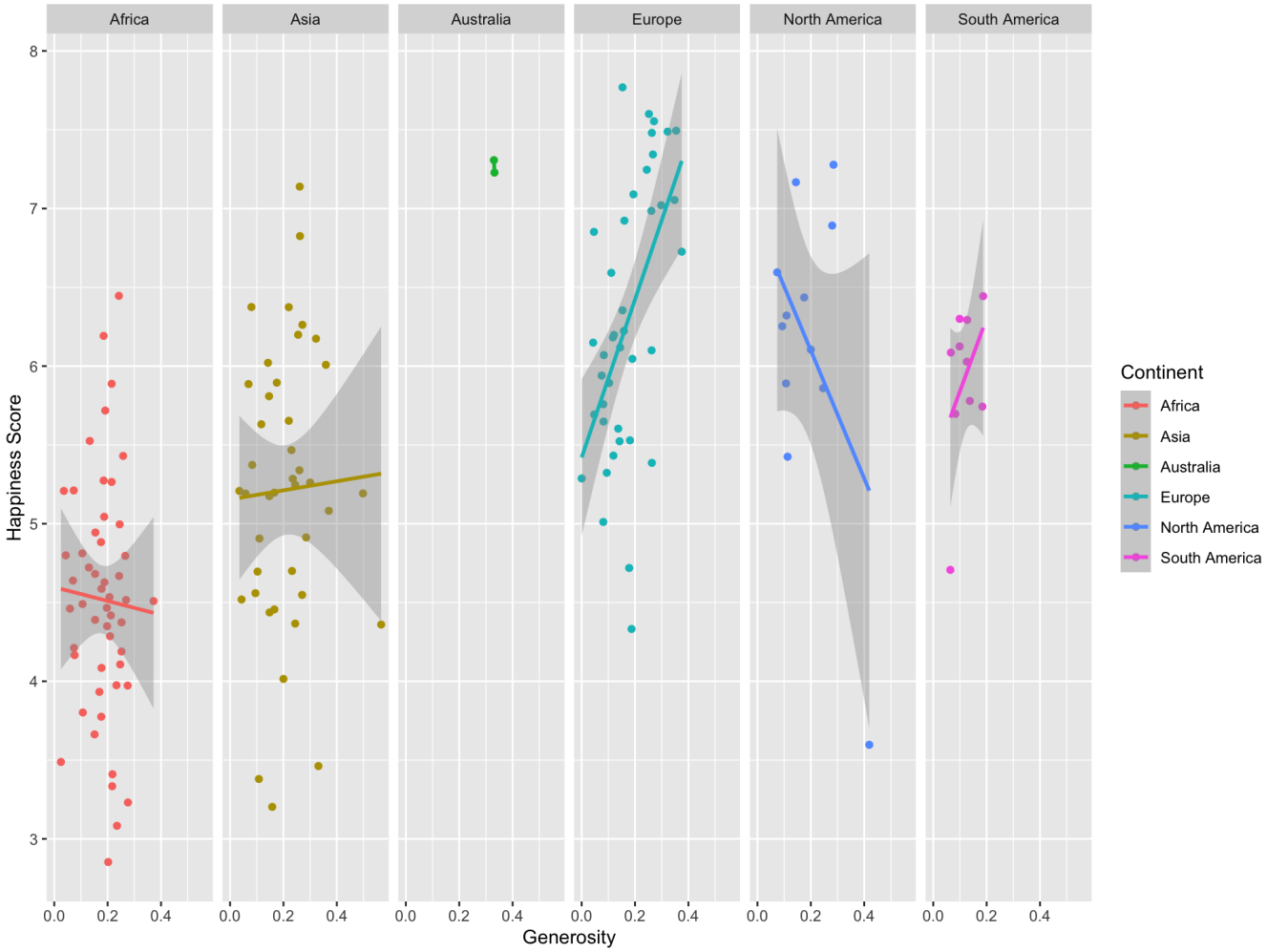


Figure 9. The relationship between Generosity and Happiness Score for different continents

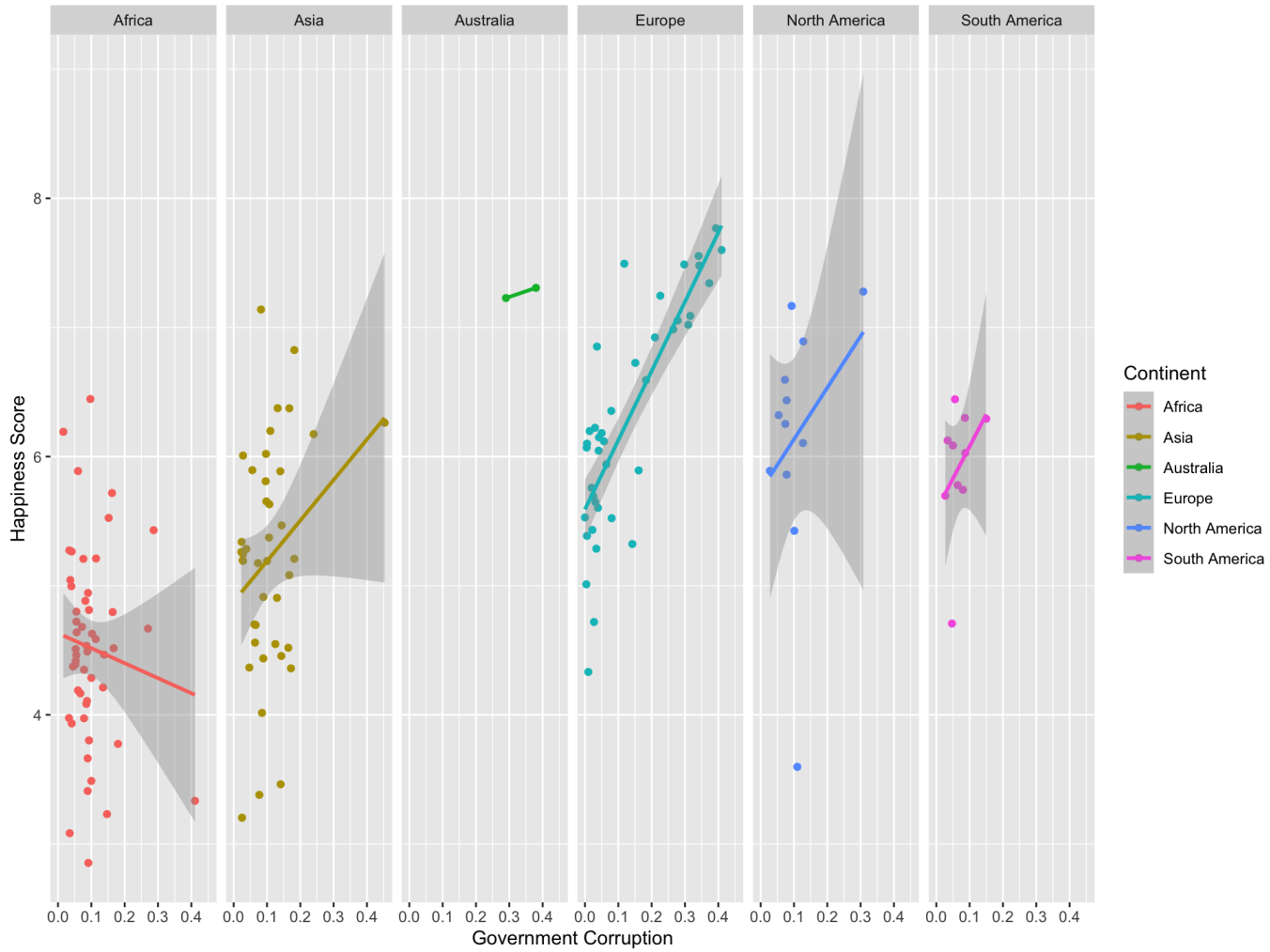


Figure 10. The relationship between Government Corruption and Happiness Score for different continents

I want to explore more about the relationship between happiness score and Generosity & Government Corruption. I am using linear regression, grouped by each continent. Based on the plots (Figure 9), the relationships between happiness and generosity are not consistent for each continent. So generosity is not a good predictor for happiness score.
Besides, from Figure 10, there is a positive correlation between corruption and happiness for most points. However, it is the opposite in Africa. So government corruption is not a good predictor for happiness score here.

```
#   Happiness_Rank     Country Happiness_Score Continent
# 1              1     Finland           7.769    Europe
# 2              2     Denmark           7.600    Europe
# 3              3      Norway           7.554    Europe
# 4              4     Iceland           7.494    Europe
# 5              5 Netherlands           7.488    Europe

#       Happiness_Rank                Country Happiness_Score Continent
# 152              152                 Rwanda           3.334    Africa
# 153              153               Tanzania           3.231    Africa
# 154              154            Afghanistan           3.203      Asia
# 155              155 Central African Republic           3.083    Africa
# 156              156            South Sudan           2.853    Africa
```

Table 2. the top and bottom 5 Countries ranked by happiness score

I also extract and compare the top five and bottom five countries of the dataset. From Table 2, the happiest five countries according to ranked scores are all European countries. The five countries listed with the lowest happiness scores are located in Africa and Asia.

3. Machine Learning

Given EDA and data visualization analysis, finally, we come to the machine learning section. We will train two different supervised machine learning models to predict the Happiness level of each country.

By transforming the Happiness Score to a binary variable "Happiness Level", countries with a score higher than the average Happiness Score are in the High level, otherwise are classified as a Low level. Logistic Regression and Random Forest will be applied to make predictions based on the transformed dataset.

3.1 Logistic Regression Model

Let's create the first linear regression model to predict the happiness level based on multiple predictors: GDP per capita, social support, healthy life expectancy, freedom, generosity, and government corruption. As a result, only Social Support, Healthy Life Expectancy, and Freedom are found to be significant at predicting the happiness level of each country, as their p-values are less than 0.05.

Reduced Logistic Regression Model

```
# Call:
# glm(formula = Happy_Level ~ Social_Support + Healthy_Life_Expectancy +
#       Freedom, family = binomial, data = Happy)

# Deviance Residuals:
#   Min       1Q   Median       3Q      Max
# -2.2563  -0.1955  -0.0017   0.3579   3.3203

# Coefficients:
#                            Estimate Std. Error z value Pr(>|z|)
#   (Intercept)              -19.107      3.527  -5.418 6.03e-08 ***
#   Social_Support             6.262      1.873   3.343 0.000828 ***
#   Healthy_Life_Expectancy   10.691      2.737   3.905 9.41e-05 ***
#   Freedom                    7.448      2.292   3.250 0.001153 **
#  ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# (Dispersion parameter for binomial family taken to be 1)

# Null deviance: 216.24  on 155  degrees of freedom
# Residual deviance:  80.78  on 152  degrees of freedom
# AIC: 88.78

# Number of Fisher Scoring iterations: 7
```

Here, I test the reduced model with these three variables (Social Support, Healthy Life Expectancy, and Freedom). Every predictor in the reduced model is found to be significant because their p-value is less than the significance level 0.05.

The final step is to verify the model on the test data to check the correctness of the results generated by the model. This step aims to make predictions by using the model with the data for which the answers are known. Then compare the results from the model with these known values.

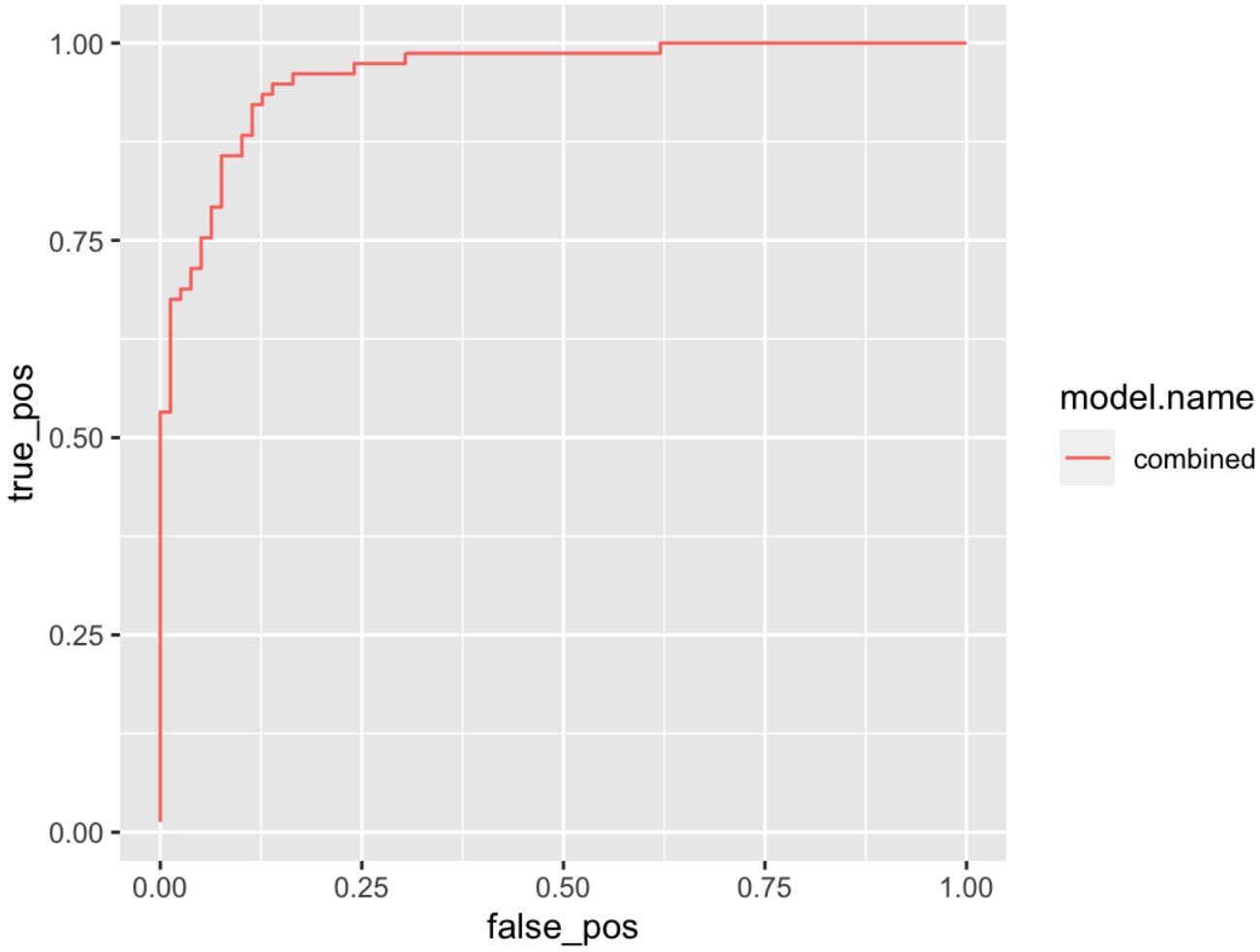


Figure 11. Display Characteristics (ROC) curve from predicted happiness level

I Split the data into a training set and test set to perform logistic regression on the training set. By comparing the result from the model with these known values., the test error of the model obtained is about 0.145. Furthermore, I calculate the area under the ROC curve (AUC) for the model, displayed in Figure 11. The higher the AUC, the better the model is at making

predictions. Thus, the model is good at predicting the happiness level of each country as the AUC score is about 0.96 (very close to 1).

3.2 Random Forest

Random Forests is a tree-based machine learning algorithm. It is also considered an "ensemble" method because it uses a set of classification trees that are calculated on random subsets of the data. Here, I also use all of the six variables in the regression.

To explore variable importance, I use the importance() function. The higher value (of importance) indicates that the variable is more important than others. By comparing the importance score, social support is the most significant factor and Government Corruption is the least important one.

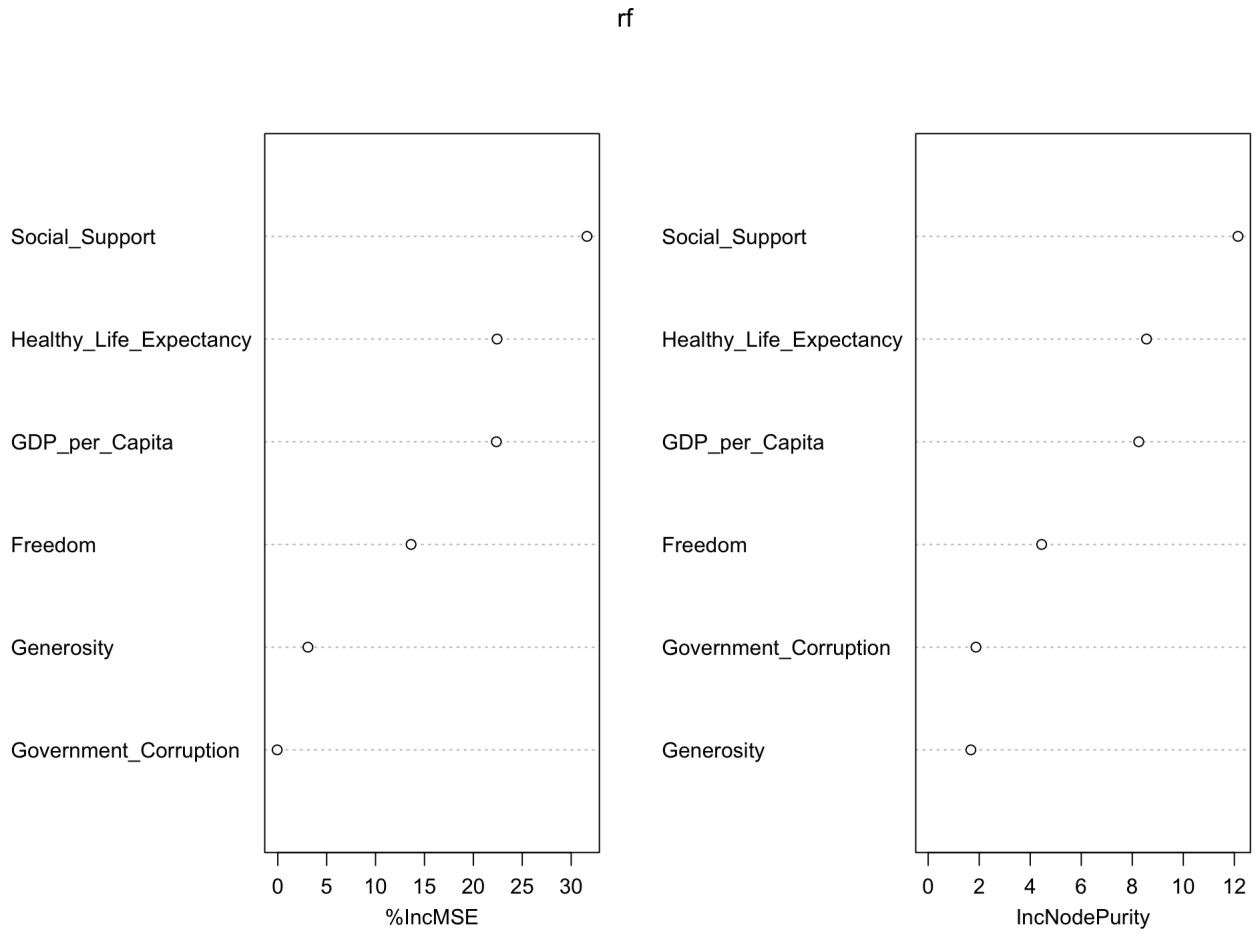


Figure 12. A dot chart of variable importance

Above I present the importance plot. The varImpPlot provides a dot chart of variable importance as measured by Random Forest - which variables are important and which are weak. From Figure 12, we see that Social support and healthy life expectancy are the most important predictors here.

4. Conclusion

After analyzing the dataset of the Global Happiness Report in 2019, I can find out the impact of the different factors in determining the Happiness Score. Now we know that GDP, Social Support, and Healthy Life Expectancy contribute the most in evaluating the happiness level of each country.

To my surprise GDP per capita is not a significant factor at predicting happiness level. By taking a closer look at the scatterplot of GDP and happiness score, I find that the economy prosperity seems to have a positive effect on happiness when it is below a certain score of GDP. Therefore, even though economic prosperity (GDP) tends to be the most responsible variable in determining happiness score, it no longer has a significant impact on happiness level while it is above a certain threshold. That is to say, money makes you happy up to a certain level, however having good health and social support is always essential for us to become happier. As the economy, physical health, and mental health are significant factors to increase happiness and well-being, governments should value both economic and socio-emotional developments to achieve national progress.

Besides, I explore the research question deeper by classifying the top 5 and bottom 5 countries in the dataset according to happiness ranking. Countries located in Europe. like Finland, Denmark, Norway, and Iceland, for their citizens, life is decent and happy. Regarding the least happy countries located in Africa and Asia, most people are unhappy perhaps because their most basic needs are far from being satisfied. This thus suggests that countries in the same continent tend to have similar happiness levels and are perhaps affected by similar factors.

Lastly, individuals should focus on both physical health and human socio-emotional development. Moreover, if governments and organizations can focus on improving these aspects, it will lead us to achieve a higher level of happiness on personal and global levels.