Manoj Munaganuru

STAT512 Project

Introduction

Aim:

To model happiness scores in every country during the year of 2020 based on several predictors using regression techniques learnt in STAT 512.

Motivation:

I believe that this study and the involved research questions could be helpful to the following groups.

1. National Policymakers: They can use this study to understand what factors would help their constituents and their fellow citizens become happier.
2. Business Investors: My study could be useful for them because multinational businesses like to put their businesses in more happy places as the happier a place is, the more likely a business will get customers. This is especially true for businesses in the hospitality and leisure sectors.
3. Vagabonds: International travelers who move to many different countries are becoming more abundant. They could use this study in order to learn what other happy countries they can go to so they can embrace the happiness in the country.

Background:

Happiness may seem like a basic word, but the effects of happiness can reverberate through all aspects of your life. With the effect that happiness has on stress, relationships, life expectancy, and health, we should not underestimate its importance. However, it is difficult to understand the most important factors that increases happiness in certain nations while leaving other nations relatively unhappy. Unhappiness is known to be at the center of different injustices like discrimination based on religion or social/political/economic affiliation which shows that the impact of unhappiness can be devastating to a nation. The only way to combat unhappiness is to improve the factors the result in greater happiness/prosperity for a nation.

The goal of this research is to understand which factors are important in evaluating the happiness of a nation, especially economic factors. These factors include ln(GDP per capita), social support, healthy life expectancy, freedom to make life choices, generosity, and perception of corruption. The significance of a factor could be important in improving happiness in a certain nation. It can also be useful to understand how a government can enact substantial policies to improve the happiness of a nation. Based on a Towards Data Science Article, a data scientist have attempted using many techniques including regression to understand happiness.

Research Questions:

1. *What predictor is the most important in predicting the mean response of Happiness score? Run diagnostics on the model to ensure no violations.*

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

2. *Which predictors in a multiple regression model without interaction terms can help predict the mean response of Happiness Score? Run relevant diagnostics and perform cross validation.*

$$H_O : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_A : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \neq 0$$

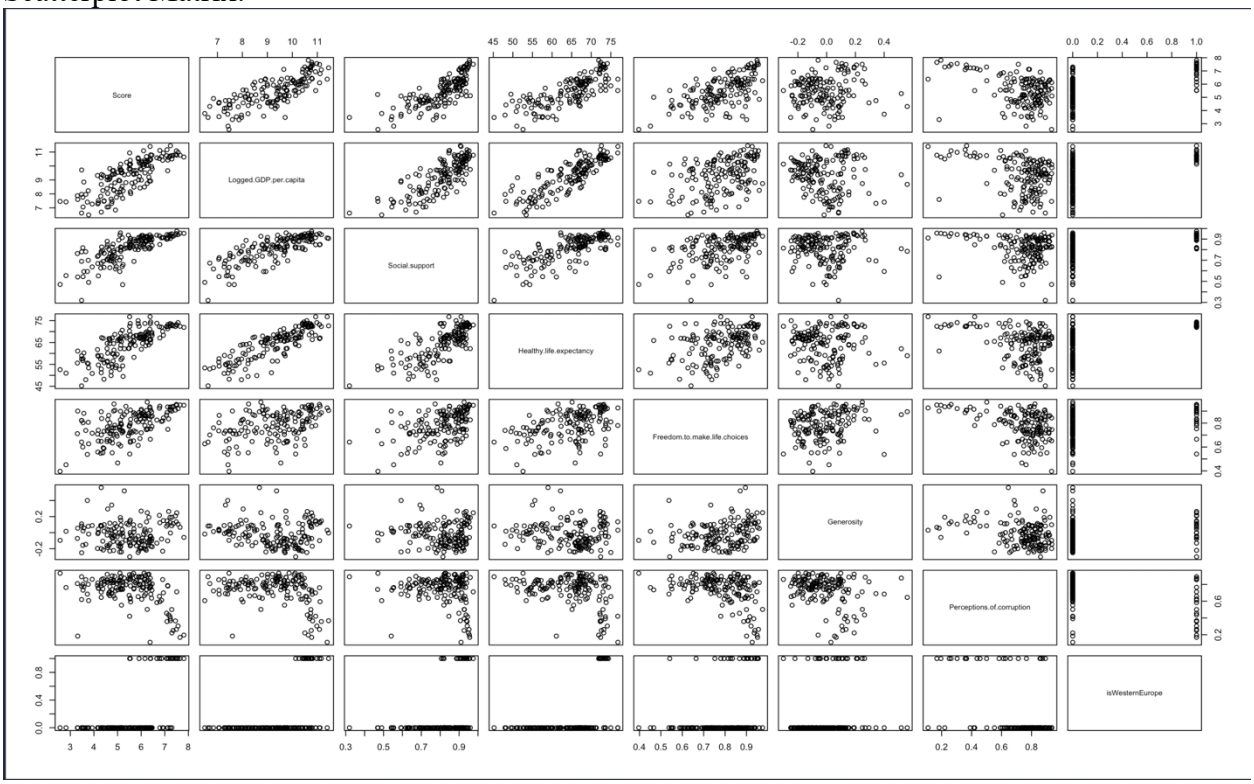
Description of the Dataset

Name	Description	Unit	Mean	Min	Max	Median
Y	Happiness	Score based on Gallup World Poll (GWP)	9.82	3.1	28.2	11.451
X_1	ln(GDP per capita)	2017 US dollar ($)	9.296	6.493	11.451	9.296
X_2	Social Support	Avg. of binary response to "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?	0.8087	0.3195	0.9747	0.8292
X_3	Healthy Life Expectancy	Years	64.45	45.20	76.80	66.31
X_4	Freedom to make life choices	Avg. of binary response to "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"	1.61	0.1	6.9	11.451
X_5	Generosity	Generosity is the residual of regressing national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.	7.4	0.4	15.2	11.451
X_6	Corruption Perception	National avg. to two binary responses in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"	12.56	1.87	17.14	11.451
X_7	isWesternEurope	Binary response to if the nation is in Western Europe	0.1373	0	1	0

This dataset has 153 rows and 9 columns. The data came from the World Happiness Report for the year 2020 and includes the above predictors/response variable. I believe it would be best to train the model on the full data rather than having a train-test data split due to the limited number of observations.
From the dataset on Kaggle, I decided to maintain the same predictors that were given in the dataset except I didn't include nation in my analysis. However, I did use region (Western Europe, Asia, etc.). The original dataset included some factors like residuals of the variables which I decided not to use as they didn't seem very helpful in creating a linear regression model to answer my research questions.

Preprocessing:
1. **isWesternEurope**: Empirically speaking, Western Europe is thought to be a region of the world with a very high level of happiness due to financial support provided as well as personal freedoms alotted. This implies that observing whether a country is or isn't in Western Europe could be a significant categorical predictor for happiness. As such, I want to make a categorical predictor called "**isWesternEurope**" where if the country's region is in Western Europe, the predictor value will be 1, else it is 0.

Preliminary Analysis
Scatterplot Matrix:



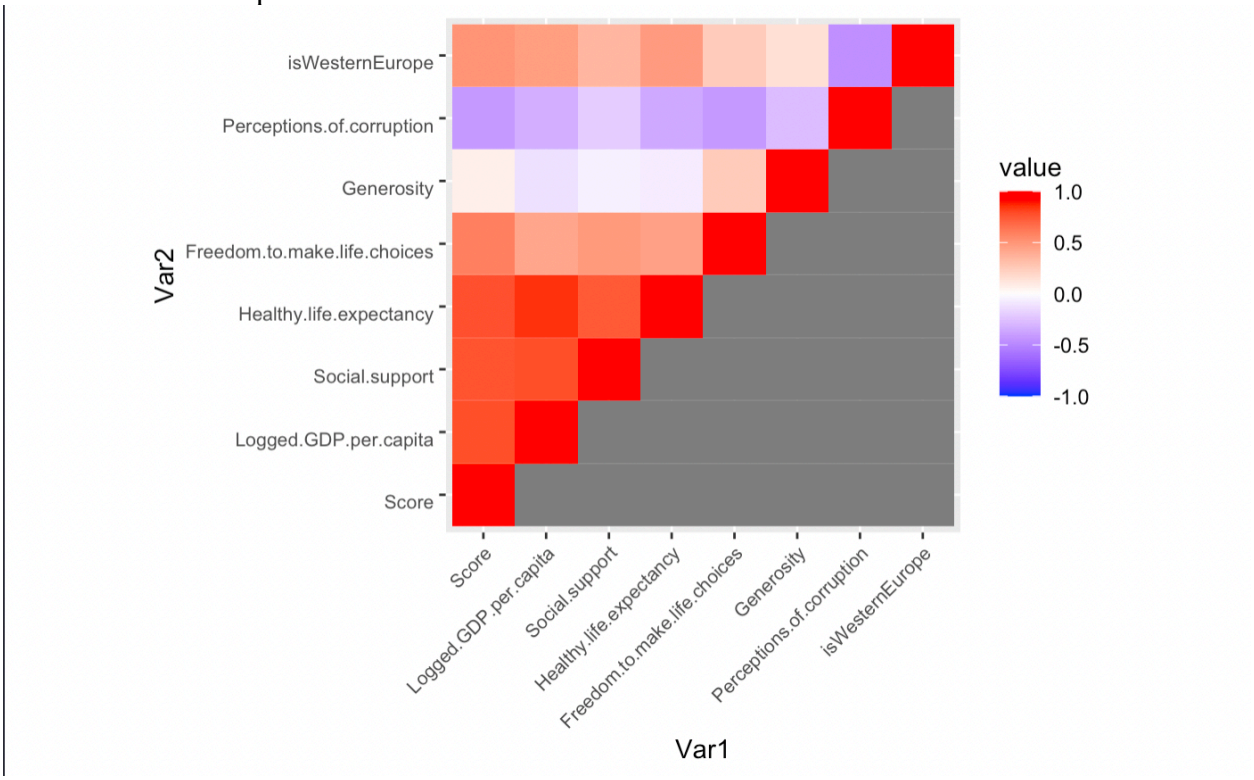
Correlation Matrix:

	Score	Logged.GDP.per.capita	Social.support	Healthy.life.expectancy
Score	1.00000000	0.7753744	0.76500076	0.77031629
Logged.GDP.per.capita	0.77537440	1.0000000	0.78181358	0.84846862
Social.support	0.76500076	0.7818136	1.00000000	0.74274409
Healthy.life.expectancy	0.77031629	0.8484686	0.74274409	1.00000000
Freedom.to.make.life.choices	0.59059678	0.4190186	0.47886318	0.44884619
Generosity	0.06904313	-0.1183994	-0.05678035	-0.07185211
Perceptions.of.corruption	-0.41830509	-0.3347291	-0.21052960	-0.35384121
isWesternEurope	0.51303795	0.4638179	0.34402632	0.47729303

	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Score	0.5905968	0.06904313	-0.4183051
Logged.GDP.per.capita	0.4190186	-0.11839937	-0.3347291
Social.support	0.4788632	-0.05678035	-0.2105296
Healthy.life.expectancy	0.4488462	-0.07185211	-0.3538412
Freedom.to.make.life.choices	1.0000000	0.25372112	-0.4201445
Generosity	0.2537211	1.00000000	-0.2784802
Perceptions.of.corruption	-0.4201445	-0.27848023	1.0000000
isWesternEurope	0.2433692	0.14100501	-0.4669186

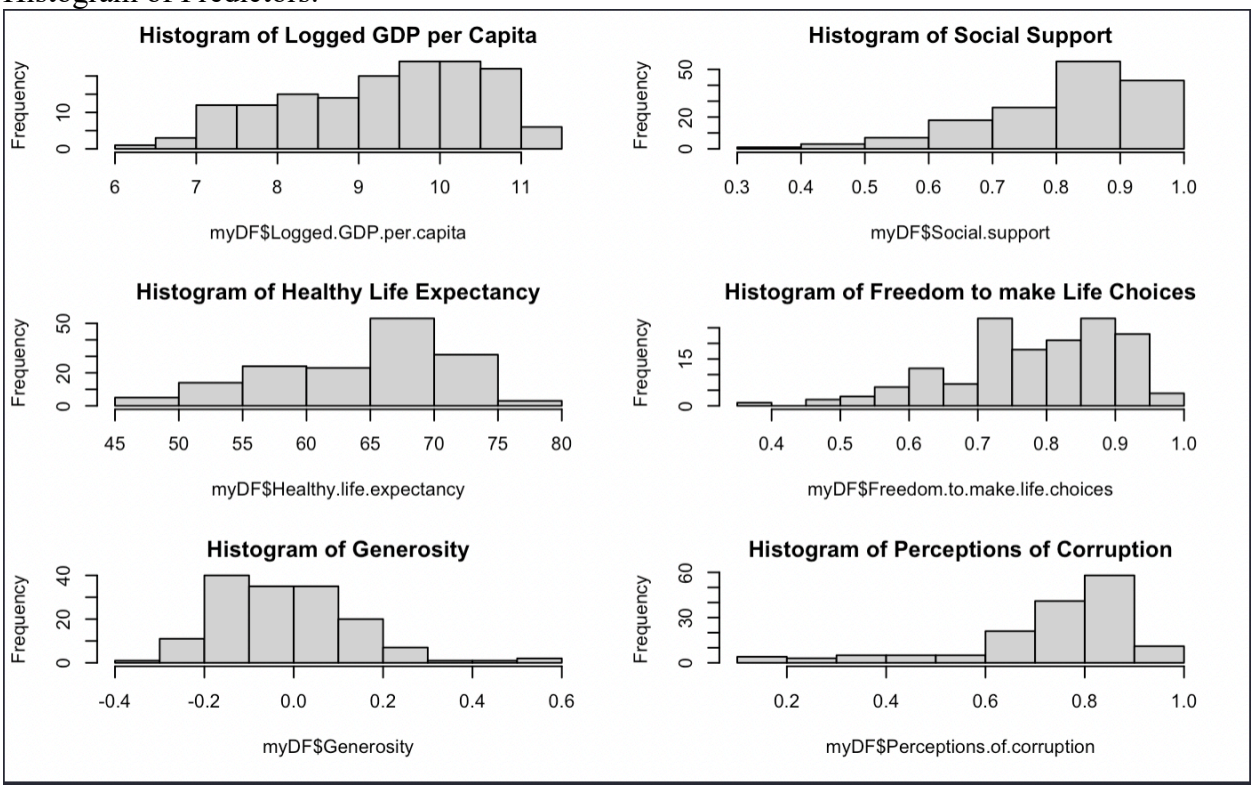
	isWesternEurope
Score	0.5130380
Logged.GDP.per.capita	0.4638179
Social.support	0.3440263
Healthy.life.expectancy	0.4772930
Freedom.to.make.life.choices	0.2433692
Generosity	0.1410050
Perceptions.of.corruption	-0.4669186
isWesternEurope	1.0000000

Correlation Heatmap:

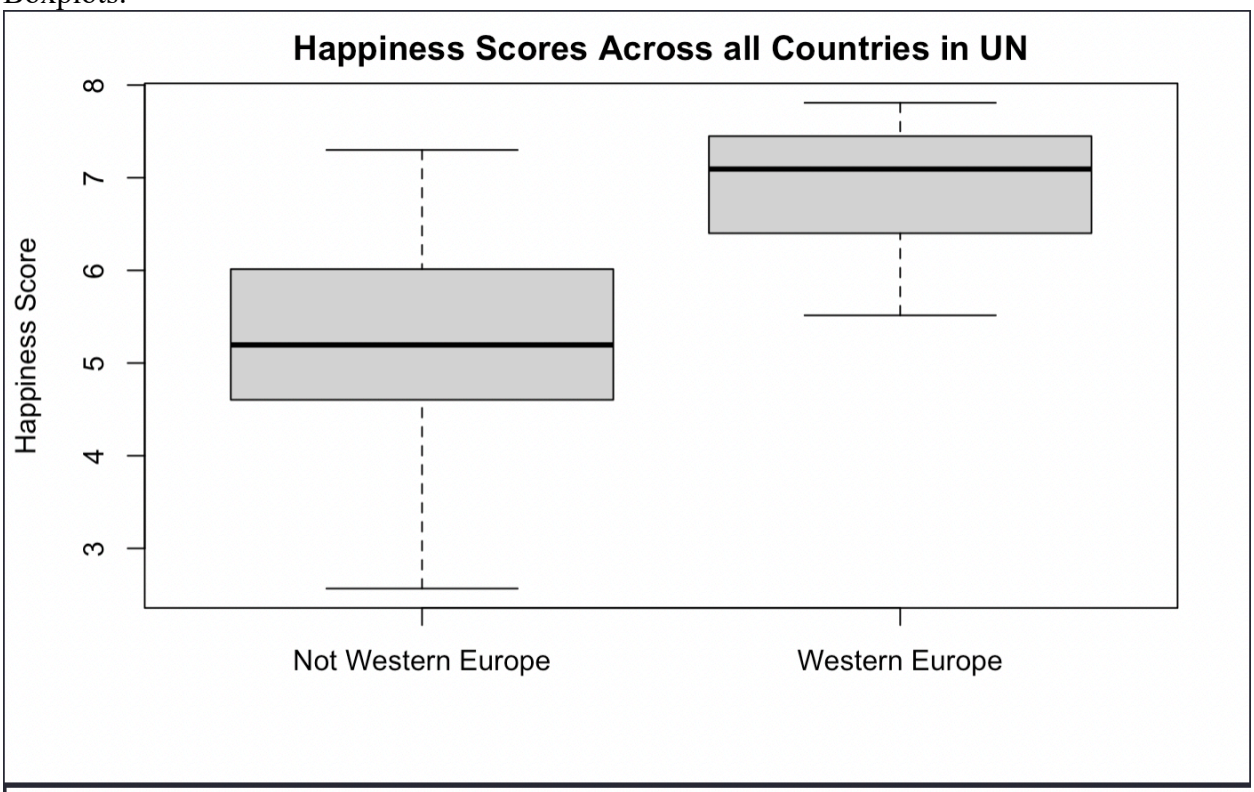


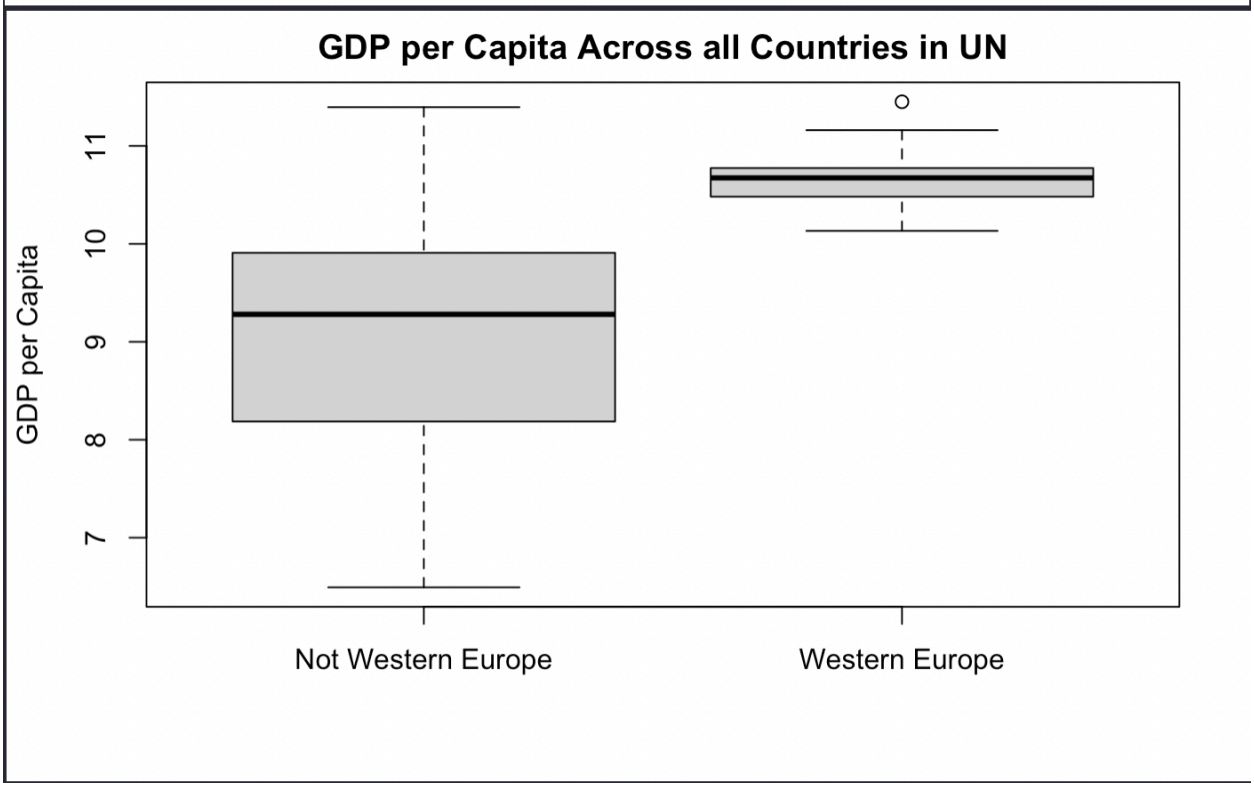
Histogram of Score:

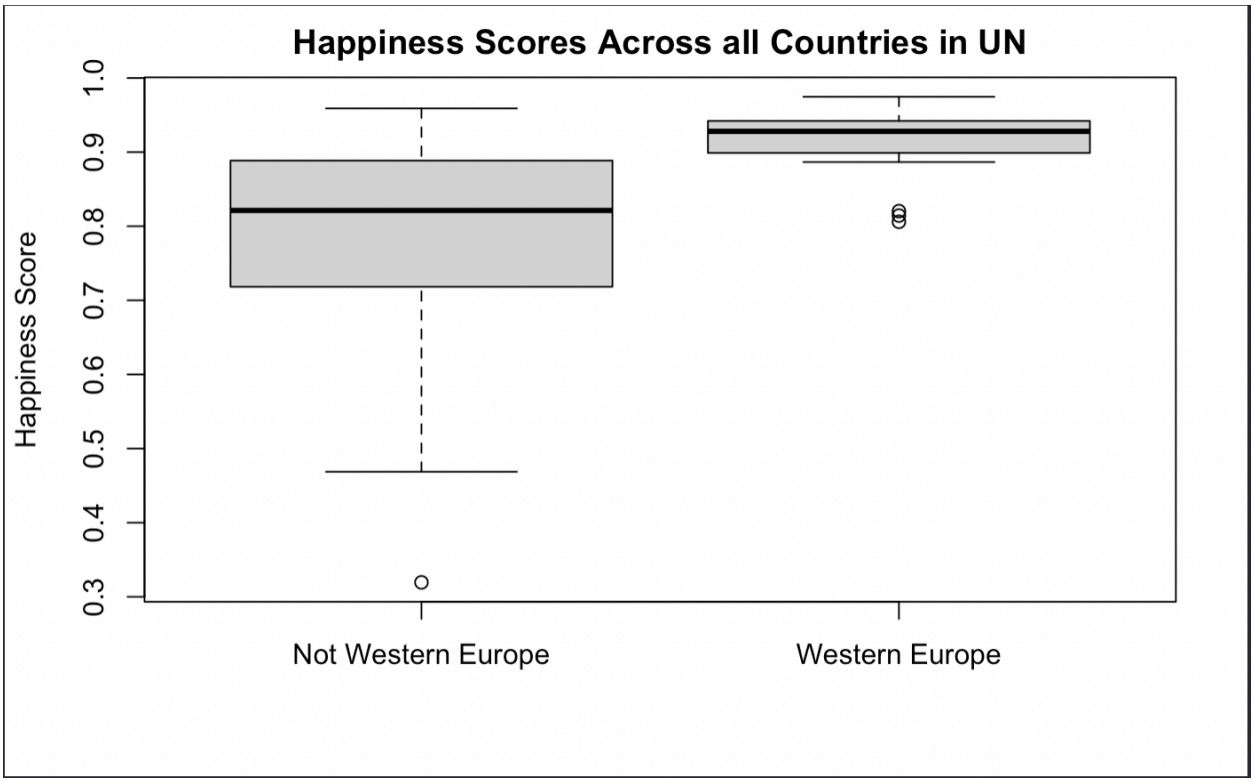
Histogram of Predictors:

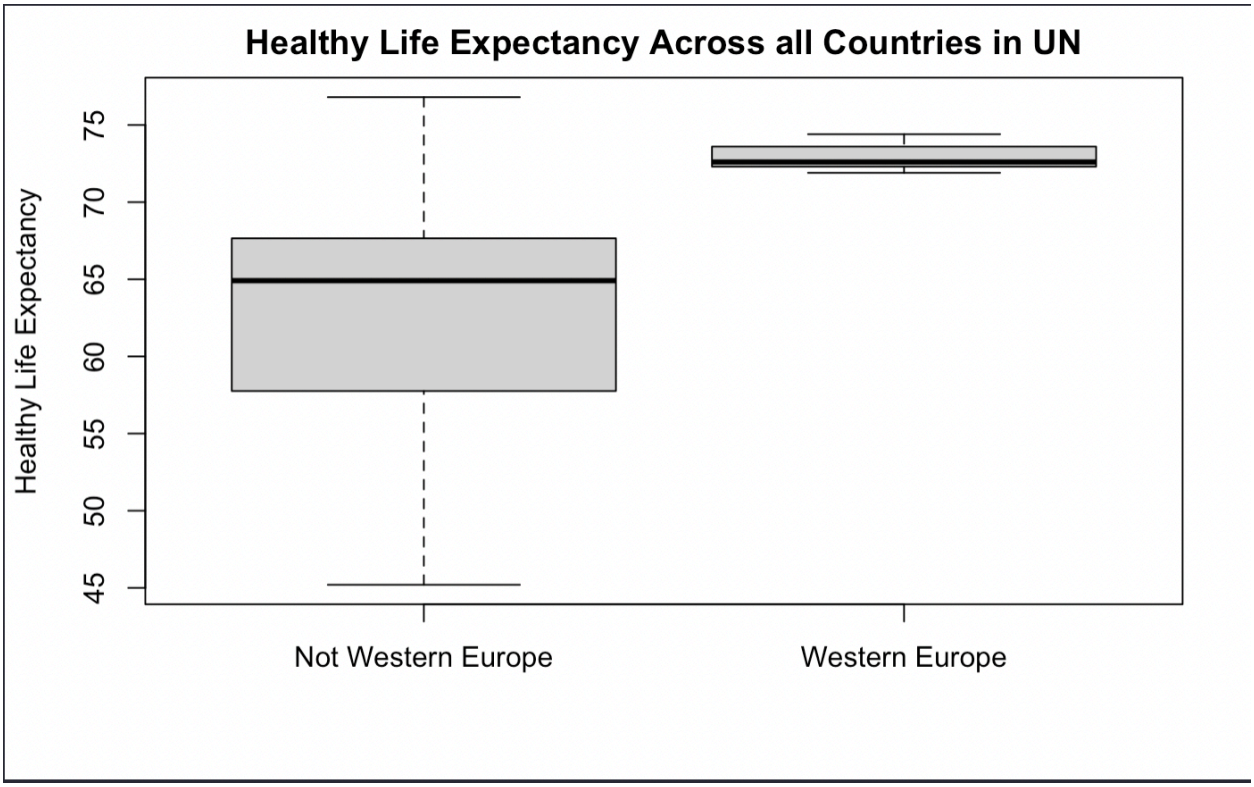


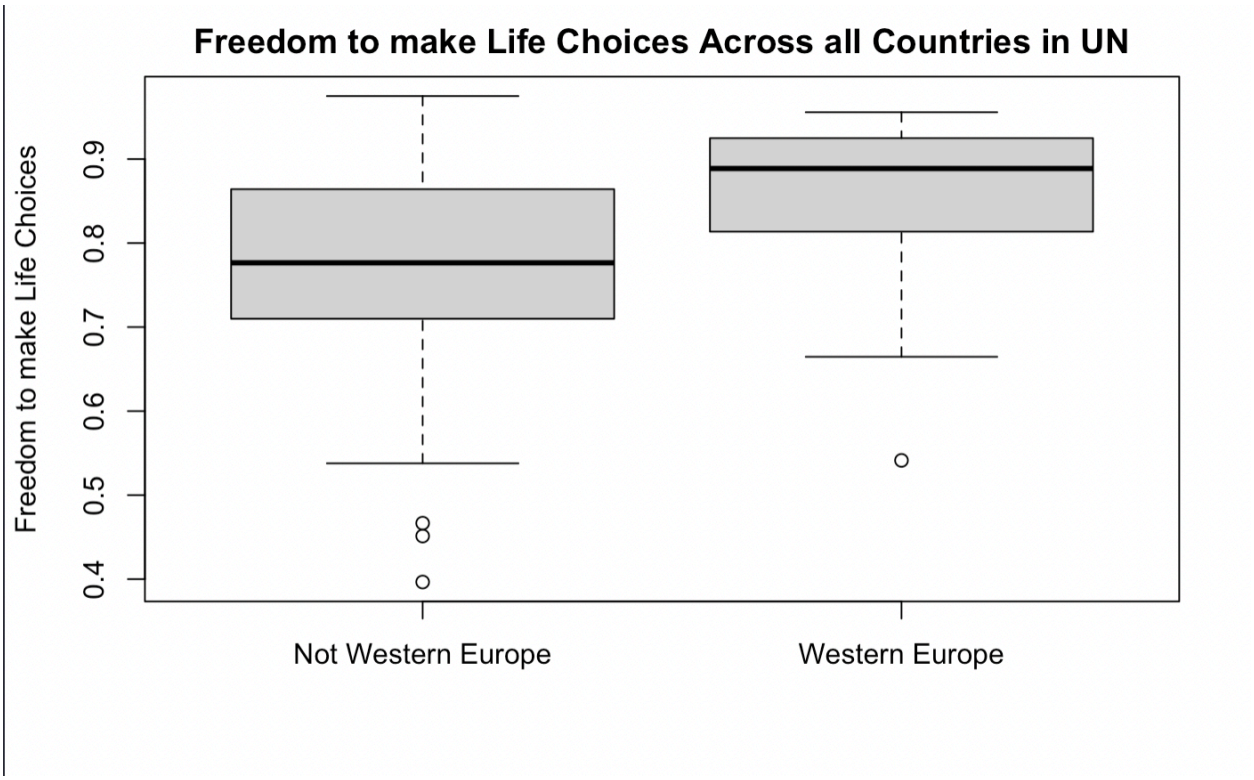
Boxplots:

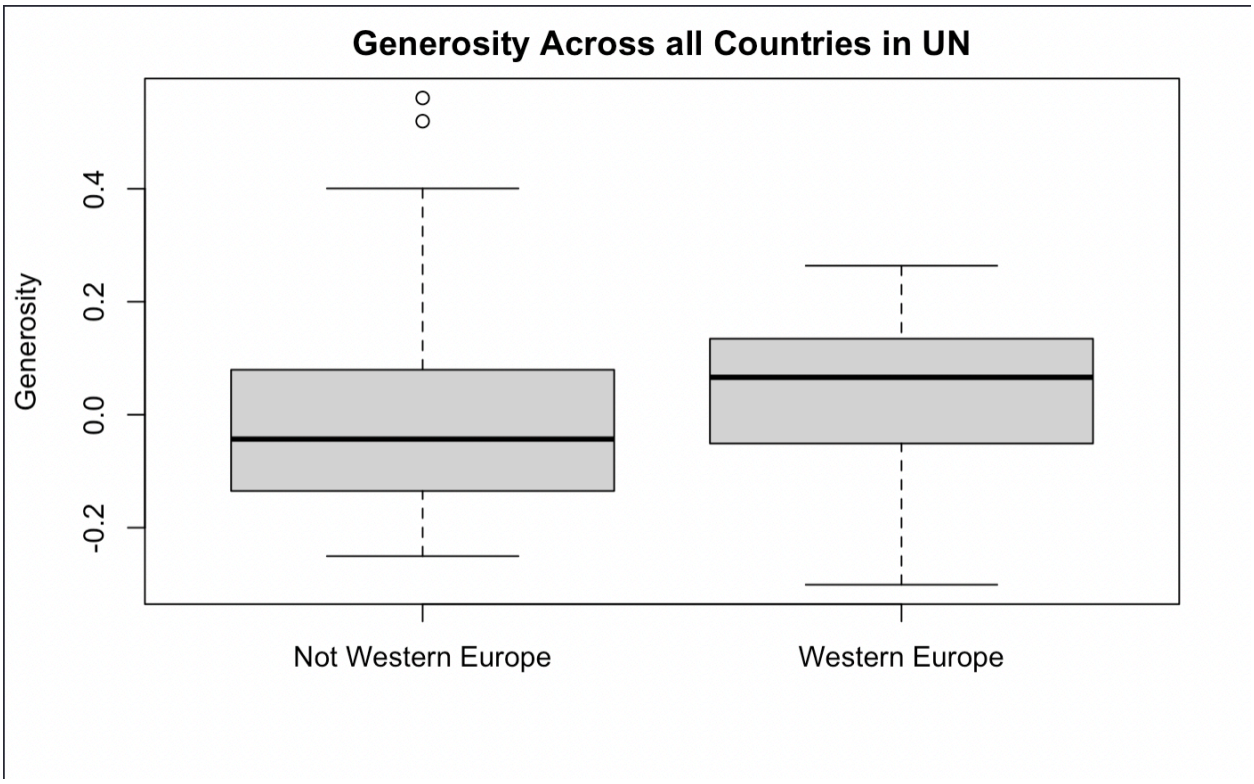

Happiness Scores Across all Countries in UN
Happiness Score
3
4
5
6
7
8
Not Western Europe
Western Europe

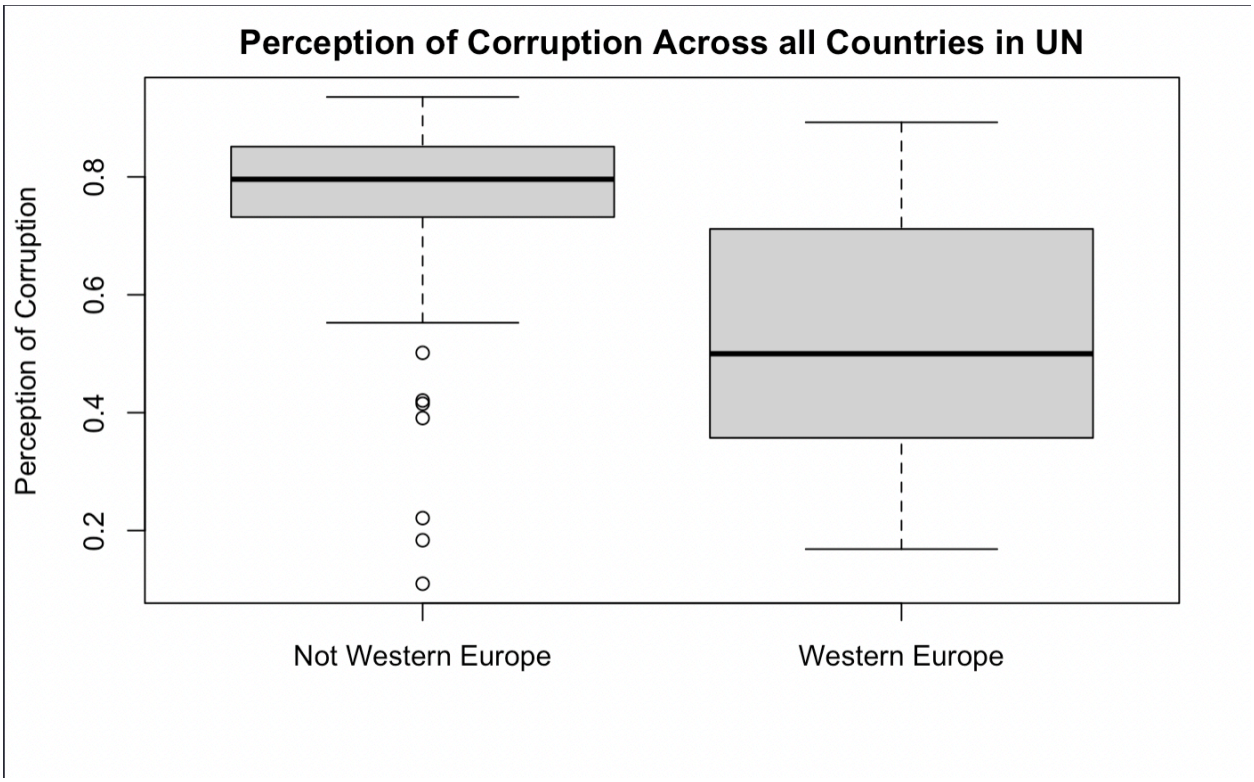

GDP per Capita Across all Countries in UN
GDP per Capita
7
8
9
10
11
Not Western Europe
Western Europe


Happiness Scores Across all Countries in UN
Happiness Score
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Not Western Europe
Western Europe


Healthy Life Expectancy Across all Countries in UN
Healthy Life Expectancy
45
50
55
60
65
70
75
Not Western Europe
Western Europe


Freedom to make Life Choices Across all Countries in UN
Freedom to make Life Choices
0.4
0.5
0.6
0.7
0.8
0.9
Not Western Europe
Western Europe


Generosity Across all Countries in UN
Generosity
-0.2
0.0
0.2
0.4
Not Western Europe
Western Europe


Perception of Corruption Across all Countries in UN
Perception of Corruption
0.2
0.4
0.6
0.8
Not Western Europe
Western Europe

Insights from the Exploratory Analysis:

1. From the scatter plot matrix, there seems to be a non-linear relationship between Y & X_5 and Y & X_6.

2. The scatter plots suggest there is a level of multicollinearity between a few predictors including but not limited to X_1 & X_2, X_1 & X_3, and X_2 & X_3. However, it is not clear how much the multicollinearity would negatively impact an MLR model.

3. Based on the histograms, most predictors have a bellshaped curve but may be slightly skewed to some direction. An exception to this includes X_4 and Y which seem to be more bimodal. This implies that the normality assumption for the residuals may be violated creating a necessity for transformations.

4. The boxplots show the significance of the X_7 due to difference in distributions of the other predictors/response that we can see in the boxplots.

5. log_{10}(GDP per capita) seems to be a good indicator of happiness score as the more wealth a person has, the happier the country. This could be concluded from the scatter plots which show that there is a relationship between Y and X_1 as well as the boxplots that show a significant difference in distribution of happiness between wealthy western European nations and the rest of the world.

6. It seems as though most of the different predictors have an impact on Y but the non-linear relationship between Y & X_5 and Y & X_6 indicates that these variables might not be included in a model or may need interaction terms/transformations.

Model Building, Diagnosis, and Validation

1. *What predictor is the most important in predicting the mean response of Happiness score? Run diagnostics on the model to ensure no violations.*

To answer this research question, I will be taking a look at the preliminary analysis to see which variables seem to have the strongest correlation with happiness score. Based on the scatterplots and correlation matrix/heatmap, $X_1, X_2, X_3, \& X_4$ seem to have the largest unique effect on the response variable. As such, those will be the variables that I will research using their respective linear models.

```r
summary(lm("Score ~ Logged.GDP.per.capita", data=myDF))
```

```
Call:
lm(formula = "Score ~ Logged.GDP.per.capita", data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29256 -0.52524  0.02843  0.57109  1.38802

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.19865    0.44586  -2.688  0.00799 **
Logged.GDP.per.capita  0.71774    0.04757  15.088  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7047 on 151 degrees of freedom
Multiple R-squared:  0.6012,    Adjusted R-squared:  0.5986
F-statistic: 227.6 on 1 and 151 DF,  p-value: < 2.2e-16
```

```r
summary(lm("Score ~ Healthy.life.expectancy", data=myDF))
```

```
Call:
lm(formula = "Score ~ Healthy.life.expectancy", data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75466 -0.58222  0.09589  0.56470  1.57394

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.350239   0.530158  -4.433 1.78e-05 ***
Healthy.life.expectancy  0.121397   0.008178  14.845  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7116 on 151 degrees of freedom
Multiple R-squared:  0.5934,    Adjusted R-squared:  0.5907
F-statistic: 220.4 on 1 and 151 DF,  p-value: < 2.2e-16
```

```r
summary(lm("Score ~ Social.support", data=myDF))
```

```
Call:
lm(formula = "Score ~ Social.support", data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01071 -0.38261 -0.04146  0.46455  2.12511

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.1926     0.3925  -0.491    0.624
Social.support  7.0059     0.4800  14.596   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7187 on 151 degrees of freedom
Multiple R-squared:  0.5852,    Adjusted R-squared:  0.5825
F-statistic: 213.1 on 1 and 151 DF,  p-value: < 2.2e-16
```

```r
summary(lm("Score ~ Freedom.to.make.life.choices", data=myDF))
```

```
Call:
lm(formula = "Score ~ Freedom.to.make.life.choices", data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-2.81474 -0.54132  0.03592  0.62210  1.85491

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.1044     0.4912   2.248    0.026 *
Freedom.to.make.life.choices 5.5771     0.6201   8.993 9.34e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9005 on 151 degrees of freedom
Multiple R-squared:  0.3488,    Adjusted R-squared:  0.3445
F-statistic: 80.88 on 1 and 151 DF,  p-value: 9.343e-16
```

Based on the summary of the models where the happiness score was regressed against the aforementioned variables, I found that all the variables except for X_4 had a significant relationship with Y. X_1 explains the most variance in happiness score with an $R^2_{adj} = 59.86\%$. Thus, I can define my hypotheses as...

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Based on the linear model summary for $y \sim \beta_1 X_1 + \beta_0 + \epsilon_i$, the p-value is $2.2 * 10^{-16}$ so we reject the null hypothesis and say that β_1 is significantly different than 0.

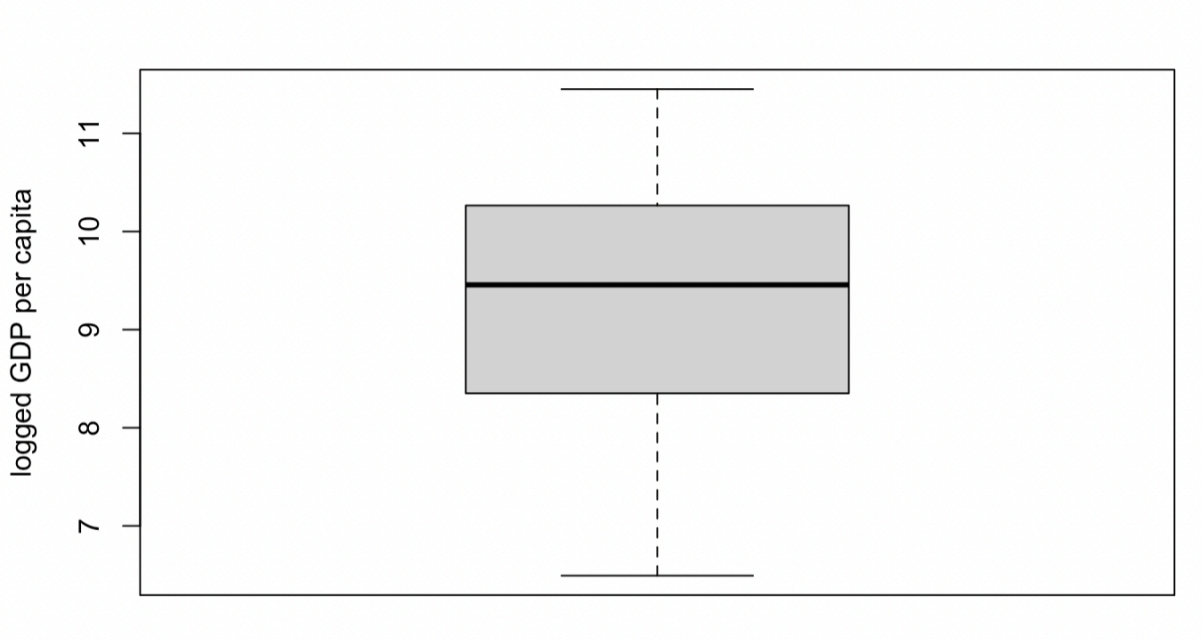
However, I still need to run diagnostics to check assumptions like (lack of) fit as well as non-constant residual variances and non-normal residuals. We will check lack of fit by using the boxplot of the predictor as well as running a lack of fit test. For testing non-constant residual variances, I will look at a residual plot vs. the predictor and perform a Brown-Forsythe Test for heteroscedasticity. In addition, to test for non-normal residuals, I will look at a QQ-Plot and perform a Shapiro-Wilk Normality Test.

Diagnostics:

Does the model have a lack of fit?

$H_O: Y_{ij} = \beta_0 + \beta_1 * X_1 + \varepsilon_{ij}$

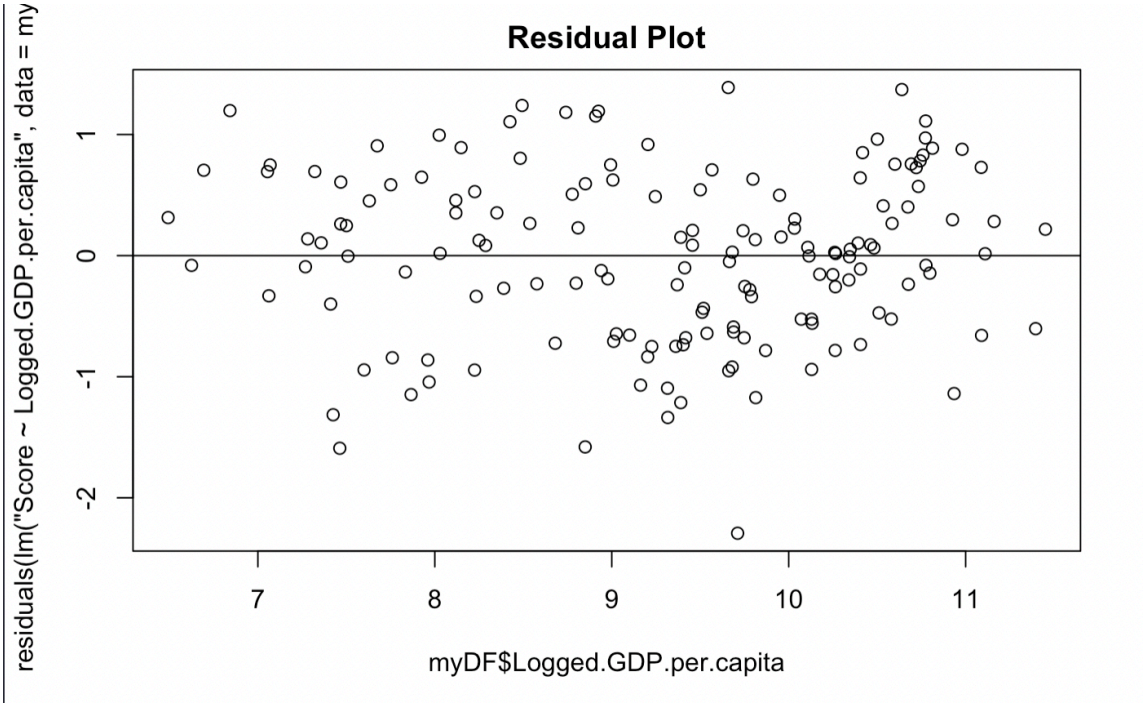
$H_A: Y_{ij} = \mu_j + \varepsilon_{ij}$




```
Analysis of Variance Table

Model 1: Score ~ Logged.GDP.per.capita
Model 2: Score ~ as.factor(Logged.GDP.per.capita)
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    151 74.992
2      1  0.194 150    74.797 2.5654 0.4666
```

We see there are no outliers in X_1, so there aren't any points on the high end that could influence the parameter estimates. We fail to reject the H_O of the lack of fit test and find that at $\alpha = 0.05$, we can say with 95% confidence that there is no lack of fit in the linear model that uses X_1 as a predictor variable for Y.

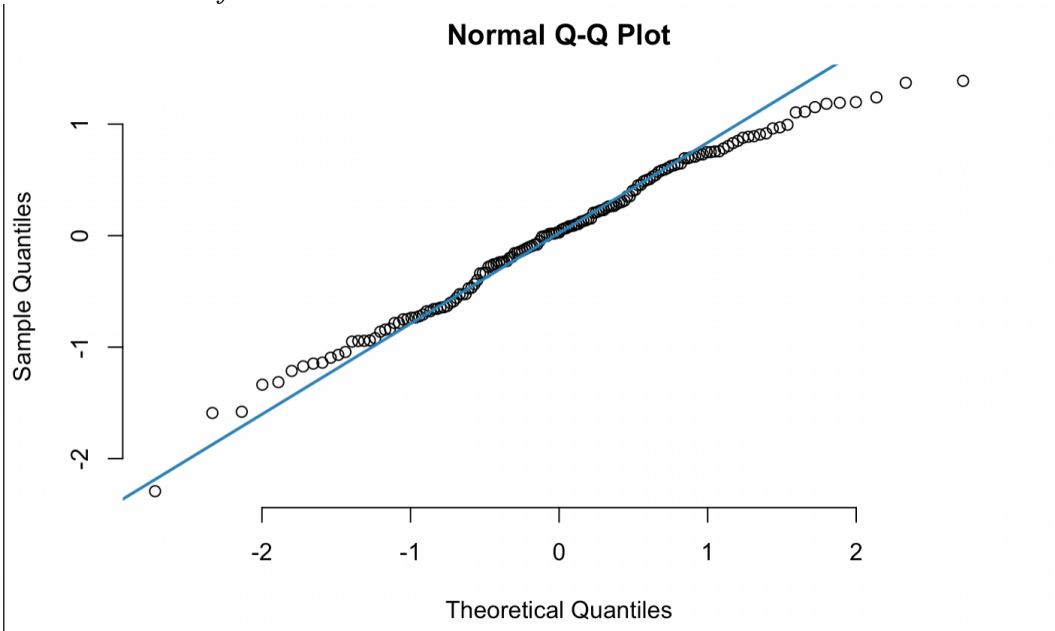
Do the residuals have non-constant variances?




```
  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : resid and group

  statistic  : 1.403352
  num df     : 2
  denom df   : 115.1583
  p.value    : 0.2499429

  Result     : Difference is not statistically significant.
-------------------------------------------------------------
```

Based on the above residual plot, there seems to be no pattern for the model residuals vs. X_1. This is confirmed by a Brown-Forsythe Test, where we fail to reject the null hypothesis at $\alpha = 0.05$ and can say with 95% confidence that the residuals have a constant variance.

Do the residuals follow a normal distribution?




```
        Shapiro-Wilk normality test

data:  residuals(lm("Score ~ Logged.GDP.per.capita", data = myDF))
W = 0.98597, p-value = 0.1249
```

Based on the above Q-Q plot, it seems that the residuals are somewhat normally distributed. After performing a Shapiro-Wilk test, we fail to reject the null hypothesis at $\alpha = 0.05$ and say with 95% confidence that the residuals are normally distributed.

Since the linear model of $Y \sim X_1$ doesn't seem to have any violations based on the plots and diagnostic tests, we need not perform any y-transformation like Box-Cox. However, the R^2_{adj} is 59.86% which shows that even though a significant amount of variance in Y is explained by X_1, adding more variables and creating an MLR model may explain more variance in Y.

As such, we can conclude that because the $Y \sim X_1$ model has an adjusted R^2_{adj} of 59.86%, X_1 seems to be the most important predictor of Happiness Score.

2. *Which predictors in a multiple regression model without interaction terms can help predict the mean response of Happiness Score? Run relevant diagnostics and perform cross validation.*

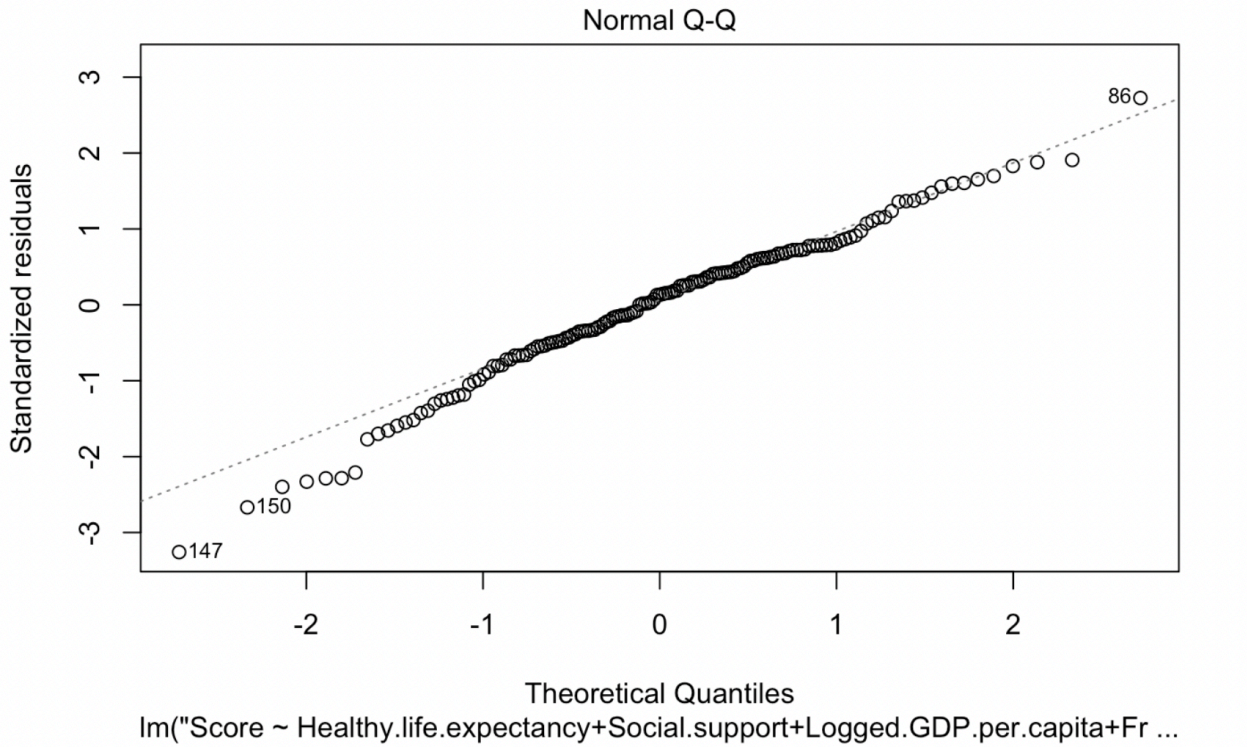
$$H_O : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_A : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \neq 0$$

First, I will build a full model with no interaction terms to get a better idea of the variance in happiness score explained by the predictors in the data.

```r
summary(lm("Score ~ Healthy.life.expectancy+Social.support+Logged.GDP.per.capita+Freedom.to.make.life.choices+Generosity
+Perceptions.of.corruption+isWesternEurope",data=myDF))
```

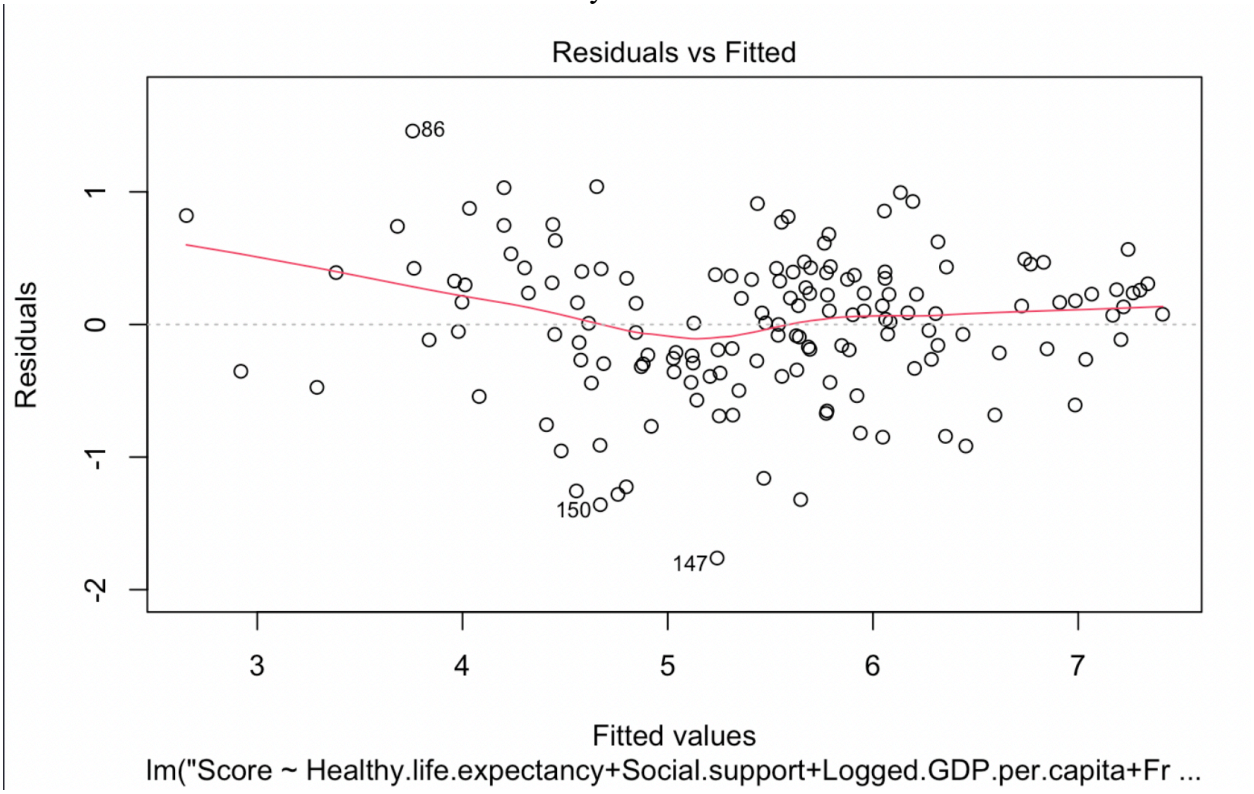
```
Call:
lm(formula = "Score ~
Healthy.life.expectancy+Social.support+Logged.GDP.per.capita+Freedom.to.make.life.choices+Generosity+Perceptions.of.corr

    data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76154 -0.29562  0.07309  0.36583  1.45880

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.85545    0.63054  -2.943 0.003790 **
Healthy.life.expectancy     0.02981    0.01285   2.319 0.021776 *
Social.support              2.75018    0.64706   4.250  3.8e-05 ***
Logged.GDP.per.capita       0.20040    0.08100   2.474 0.014517 *
Freedom.to.make.life.choices  1.93487    0.49027   3.947 0.000123 ***
Generosity                  0.28374    0.33305   0.852 0.395650
Perceptions.of.corruption  -0.34409    0.32510  -1.058 0.291628
isWesternEurope             0.44587    0.16309   2.734 0.007041 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5571 on 145 degrees of freedom
Multiple R-squared:  0.7607,    Adjusted R-squared:  0.7491
F-statistic: 65.84 on 7 and 145 DF,  p-value: < 2.2e-16
```

This model seems to have a good R^2_{adj} but we need to determine if the model has any violations by running diagnostics.



```
        Shapiro-Wilk normality test

data:  lm("Score ~
Healthy.life.expectancy+Social.support+Logged.GDP.per.capita+Freedom.to.make.life.choices+Generos
 data = myDF)$residuals
W = 0.98321, p-value = 0.05955
```

The residuals seem to be normally distributed based on the QQ-plot above. This is confirmed by the Shapiro-Wilk test where we fail to reject the null hypothesis at 0.05 alpha level and say with 95% confidence that the residuals are normally distributed.




```
  Brown-Forsythe Test (alpha = 0.05)
------------------------------------------------------------
  data : residual and group

  statistic   : 3.758735
  num df      : 4
  denom df    : 26.78383
  p.value     : 0.01490096

  Result      : Difference is statistically significant.
------------------------------------------------------------
```

However, we reject the null hypothesis of the Brown-Forsythe Test at $\alpha = 0.05$ and say with 95% confidence that the residuals don't have constant variance. Since the diagnostics for the full model failed, we should take a look at best subset algorithm to get a better MLR model using criterions such as $SSE_p, AIC, PRESS_p, SBC, C_p, \& R^2_{adj}$.

```{r}
library(ALSM)
BestSub(myDF[,4:10], myDF$Score, num=1)
```

```
Loading required package: leaps
Loading required package: SuppDists
Loading required package: car
Loading required package: carData
  p 1 2 3 4 5 6 7     SSEp        r2    r2.adj        Cp      AICp       SBCp   PRESSp
1 2 1 0 0 0 0 0 0 74.99173 0.6012055 0.5985644 92.610995 -105.0982  -99.03732 76.91505
2 3 1 0 0 1 0 0 0 58.88904 0.6868371 0.6826616 42.730779 -140.0818 -130.99048 61.09347
3 4 0 1 1 1 0 0 0 52.54787 0.7205585 0.7149321 24.300578 -155.5131 -143.39139 56.02698
4 5 1 1 0 1 0 0 1 47.43941 0.7477245 0.7409062  9.841967 -169.1606 -154.00845 50.57271
5 6 1 1 1 1 0 0 1 45.71769 0.7568803 0.7486110  6.294859 -172.8168 -154.63413 49.76630
6 7 1 1 1 1 0 1 1 45.23068 0.7594701 0.7495854  6.725810 -172.4553 -151.24225 50.21860
7 8 1 1 1 1 1 1 1 45.00541 0.7606681 0.7491142  8.000000 -171.2193 -146.97576 50.81566
```

When running Best Subset Algorithm, looking at the predictors using Happiness Score as the response variable, we see the model with all variables except "Generosity" and "Perceptions.of.corruption" seems to have the best values for C_p, AIC_p, SBC_p, and $PRESS_p$ while the model with all variables except "Generosity" has the best values for R^2_{adj}. However, SBC_p usually prefers more general models whereas AIC_p and C_p prefers more complex models and because all criterions chose the same model, it seems that $Y \sim X_1 + X_2 + X_3 + X_4 + X_7$ is the best model based on the output of the Best Subset Algorithm.

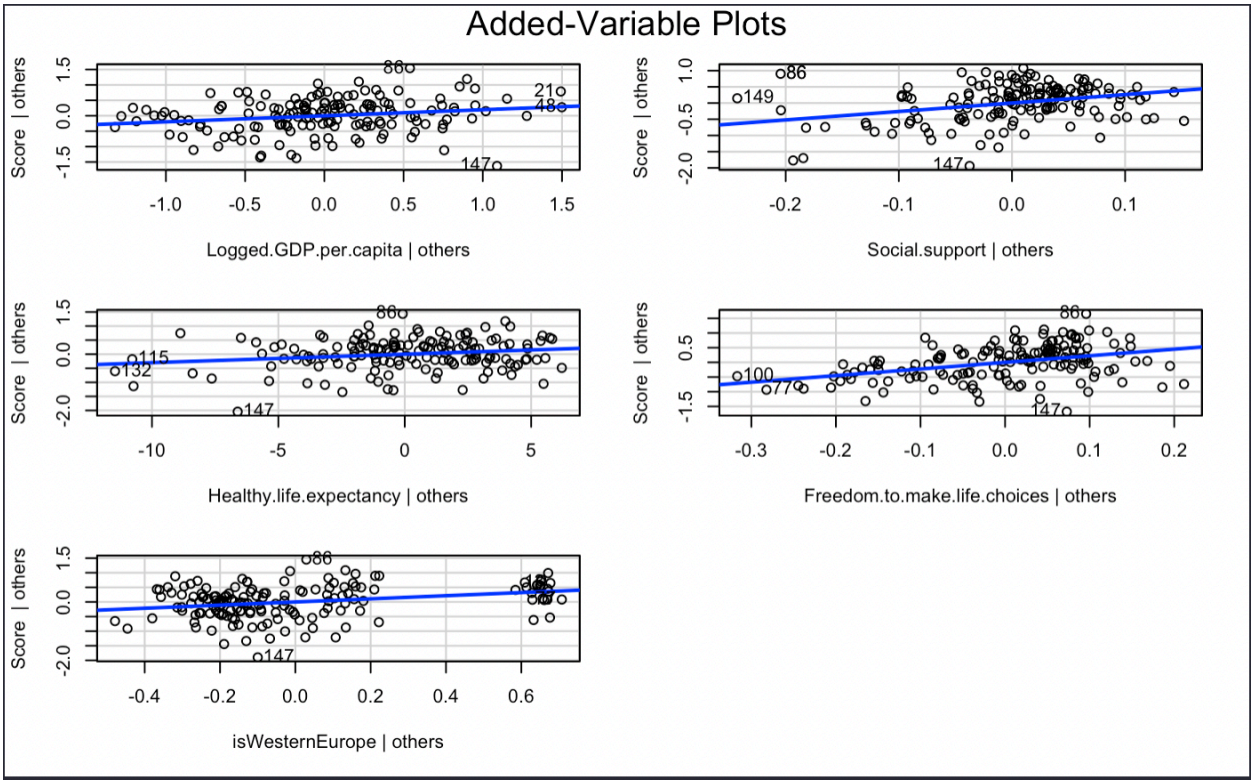
I will run diagnostics on the following model selected by Best Subset to determine if there are any MLR assumptions that are violated.

```r
MLR_no_interaction = lm("Score ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy +
Freedom.to.make.life.choices + isWesternEurope", data= myDF)
summary(MLR_no_interaction)
```

```
Call:
lm(formula = "Score ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy + Freedom.to.make.life.choices
+ isWesternEurope",
    data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84438 -0.30447  0.06413  0.34559  1.43684

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.25042    0.48640  -4.627 8.09e-06 ***
Logged.GDP.per.capita       0.19733    0.07962   2.478 0.014331 *
Social.support              2.60615    0.63511   4.103 6.72e-05 ***
Healthy.life.expectancy     0.03016    0.01282   2.353 0.019954 *
Freedom.to.make.life.choices  2.25204    0.44370   5.076 1.15e-06 ***
isWesternEurope             0.53762    0.15096   3.561 0.000498 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5577 on 147 degrees of freedom
Multiple R-squared:  0.7569,    Adjusted R-squared:  0.7486
F-statistic: 91.53 on 5 and 147 DF,  p-value: < 2.2e-16
```

Diagnostics

AV Plots:



These AV plots indicate that a linear term of each of the predictors would be a helpful addition in the regression model.

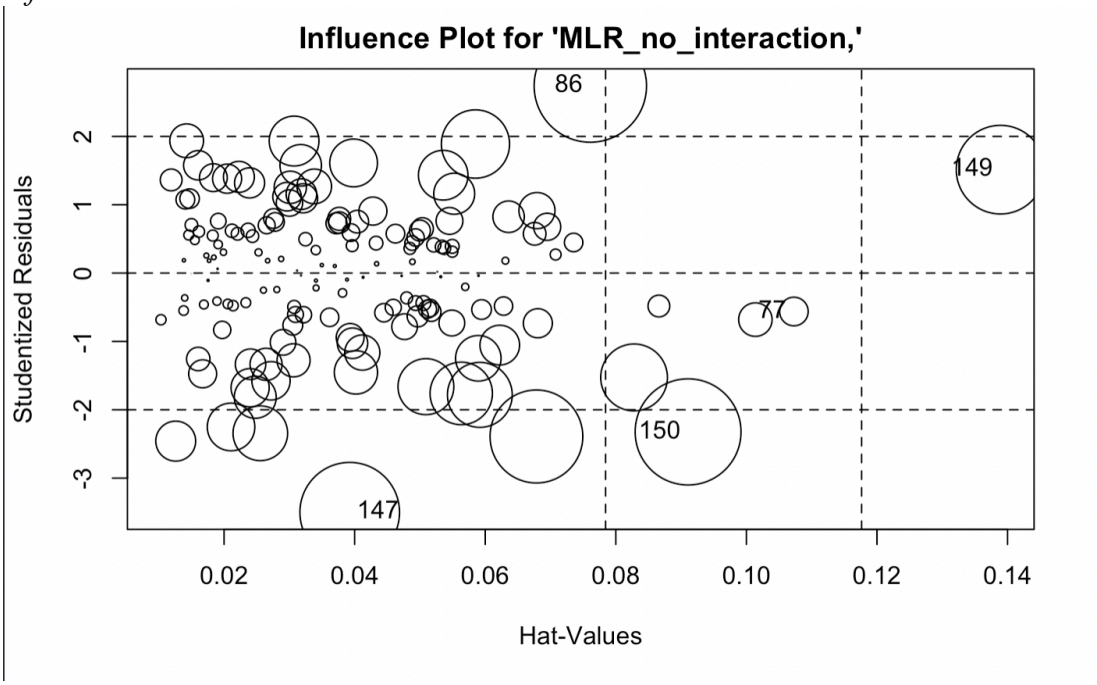
Is there multicollinearity among predictors?
Let's take a look at the VIF coefficients.

```r
library(fmsb)
VIF(lm(Logged.GDP.per.capita ~ Social.support + Healthy.life.expectancy + Freedom.to.make.life.choices + isWesternEurope, myDF))
VIF(lm(Social.support ~ Logged.GDP.per.capita + Healthy.life.expectancy + Freedom.to.make.life.choices + isWesternEurope, myDF))
VIF(lm(Healthy.life.expectancy ~ Logged.GDP.per.capita +  Social.support + Freedom.to.make.life.choices + isWesternEurope, myDF))
VIF(lm(Freedom.to.make.life.choices ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy + isWesternEurope, myDF))
VIF(lm(isWesternEurope ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy + Freedom.to.make.life.choices, myDF))
```

```
Registered S3 methods overwritten by 'fmsb':
  method    from
  print.roc pROC
  plot.roc  pROC
[1] 4.473579
[1] 2.908013
[1] 4.000525
[1] 1.334894
[1] 1.327579
```

We can see that all of the VIF values for the model are greater than one. The largest VIF value is 4.473579 Since the maximum VIF for the model is less than 10, we can conclude that there is not excessive multicollinearity between the predictors used in the model. This result was slightly surprising as there seemed to be a significant degree of multicollinearity in the data based on the scatterplots/correlation matrix (heatmap) from the **Exploratory Analysis**.

Influential Point Analysis
Influence Plot:



	StudRes <dbl>	Hat <dbl>	CookD <dbl>
77	-0.5622525	0.10727913	0.006361158
86	2.7391225	0.07609628	0.098630088
147	-3.5009291	0.03924448	0.077506274
149	1.5135777	0.13891765	0.061062378
150	-2.3238852	0.09106818	0.087559578

Are there Y outliers? What about X outliers?

```{r}
ti <- rstudent(MLR_no_interaction)
ti_test_stat <- qt(1 - 0.05/(2*nrow(myDF)), nrow(myDF) - 6 - 1)
ti[abs(ti) > ti_test_stat]
```

```
named numeric(0)
```

None of the observations are Y outliers based on the studentized-deleted residuals above.

```{r}
hii_bar <- mean(lm.influence(MLR_no_interaction)$hat, na.rm = TRUE)
lm.influence(MLR_no_interaction)$hat[lm.influence(MLR_no_interaction)$hat > 2*hii_bar]
```

```
       77        132        143        149        150        153 
0.10727913 0.08659873 0.08279591 0.13891765 0.09106818 0.10137624 
```

Based on the diagonal values of the hat matrix, the above points are the only observation values which are greater than twice mean diagonal value in the hat matrix. This mean diagonal value is called the mean leverage value.

Identifying Influential Points

This will be based on DFFITS (influence on a single fitted value), Cook's Distance (influence on all fitted values), and DFBETAS (influence on the regression coefficients).

$\frac{n}{p} > 10$ so this dataset would be considered large.

DFFITS:

```{r}
dffits(MLR_no_interaction)[abs(dffits(MLR_no_interaction)) > 2*sqrt(6/nrow(myDF))]
```

```
       76        85        86       143       144       146       147       149       150 
-0.4469558  0.4708218  0.7861036 -0.4587564 -0.6450062 -0.4301831 -0.7075647  0.6079405 -0.7355841 
```

```
As such, the above points have a heavy influence on the fitted values of the regression function.
```

Cook's Distance:

```{r}
cooks_test_f_values <- pf(cooks.distance(MLR_no_interaction), 6, nrow(myDF)-6)
length(cooks_test_f_values[cooks_test_f_values < 0.2])
length(cooks_test_f_values[cooks_test_f_values > 0.5])
nrow(myDF)
```

```
[1] 153
[1] 0
[1] 153
```

As such, since all fitted values have a value from F(p,n-p) that is less than the 20th percentile, all the points in the data have very little influence on the fitted value.

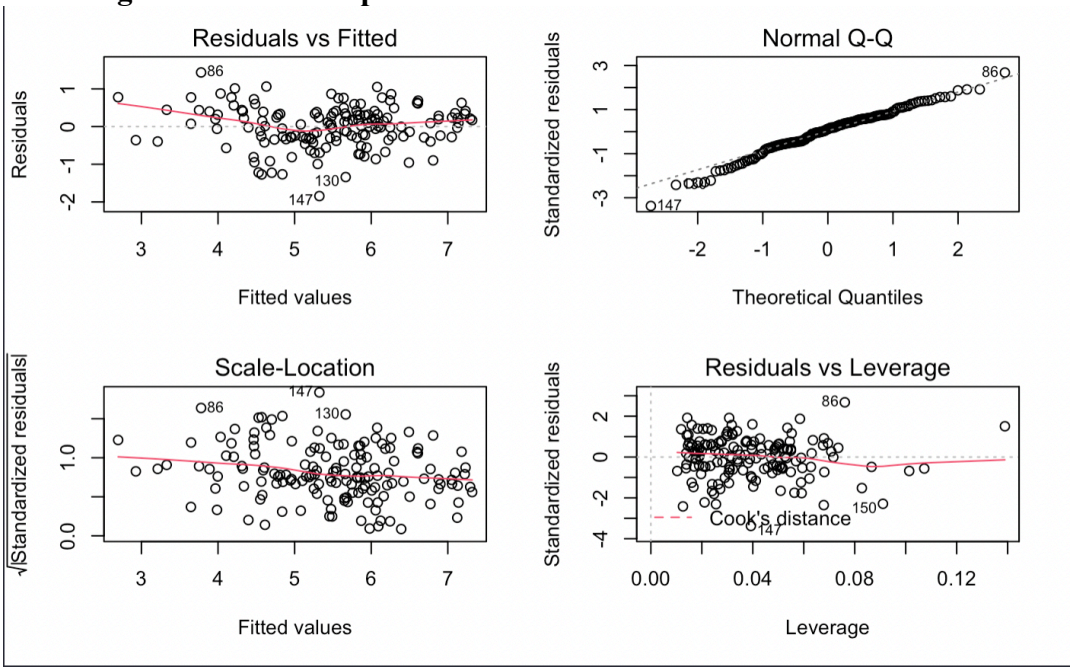
DFBETAS:

```{r}
length(dfbetas(MLR_no_interaction)[abs(dfbetas(MLR_no_interaction)) > 2])
```

```
[1] 0
```

Based on the DFBETAS, there are no points that are heavily influential on the regression coefficients. Through the influential point analysis, I didn't want to remove any influential points as it is bad practice to remove points if they result in a bad fitting model. As such, I didn't remove any observations from the data.

Checking Residual Assumptions:



Based on the plots above, there doesn't seem to be a pattern between the residuals and the fitted values. The QQ-plot also seems to show that the residuals are normally distributed.

Let's run some diagnostics to confirm the above results.

```
        Shapiro-Wilk normality test

data:  MLR_no_interaction$residuals
W = 0.98429, p-value = 0.07969


  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : residual and group

  statistic  : 1.892977
  num df     : 3
  denom df   : 41.9874
  p.value    : 0.1454631

  Result     : Difference is not statistically significant.
-------------------------------------------------------------
```

For the $Y \sim X_1 + X_2 + X_3 + X_4 + X_7$ model, the Shapiro-Wilk Test gives a p-value of 0.07969 which is greater than $\alpha = 0.05$ so we fail to reject the null hypothesis of the Shapiro-Wilk Test and say with 95% confidence that the model has residuals which follow the normal distribution. Also, the Brown-Forsythe Test gives a p-value of 0.1454631 which is greater than $\alpha = 0.05$ so we fail to reject the null hypothesis of the Brown-Forsythe Test and say that residuals have constant variances at the 95% confidence level.

Let's run a global ANOVA F-test to confirm the results of our model.

```r
anova(MLR_no_interaction)
f_stat = (113.054 + 12.198 + 4.561 + 8.571 + 3.944)/0.311
f_stat
f_stat > qf(0.95, 7-1, nrow(myDF)-7)
```

```
Analysis of Variance Table

Response: Score
                             Df  Sum Sq Mean Sq F value    Pr(>F)
Logged.GDP.per.capita         1 113.054 113.054 363.513 < 2.2e-16 ***
Social.support                1  12.198  12.198  39.221 3.946e-09 ***
Healthy.life.expectancy       1   4.561   4.561  14.665 0.0001896 ***
Freedom.to.make.life.choices  1   8.571   8.571  27.558 5.240e-07 ***
isWesternEurope               1   3.944   3.944  12.683 0.0004977 ***
Residuals                   147  45.718   0.311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 457.6463
[1] TRUE
```

EacI completed an ANOVA F-test on the predictors I used in the MLR model, I found that each predictor was significant so we can reject the null hypothesis proposed in the research question and say that at least some of the predictors are significantly different than 0.

As such, since the model doesn't have any violations and the beta parameters are significantly different than 0, we can say that $Y \sim X_1 + X_2 + X_3 + X_4 + X_7$ is the **final model** for this research question. This is because it explains a large portion of the variance in Y with an R^2_{adj}=74.86%.

However, a model with interaction terms may explain more variance so I am going to research further to see if a model with interaction terms is better at explaining the variance in Y.

MLR Model with Interaction Terms

```r
myDF$isWesternEurope_gdp = myDF$isWesternEurope * myDF$Logged.GDP.per.capita
myDF$isWesternEurope_social = myDF$isWesternEurope * myDF$Social.support
myDF$isWesternEurope_life_expectancy = myDF$isWesternEurope * myDF$Healthy.life.expectancy
myDF$isWesternEurope_freedom = myDF$isWesternEurope * myDF$Freedom.to.make.life.choices
```

I created the above 4 interaction variables in order to see if interaction terms could improve the amount of explained variance in the model but also to reduce some of the influential points that were found in the MLR model without interaction terms. I decided to do Best Subset Algorithm on all of these predictors and got the following.

```
library(ALSM)
BestSub(myDF[,c(4:14)], myDF$Score, num=1)
```

```
    p 1 2 3 4 5 6 7 8 9 A B     SSEp        r2    r2.adj        Cp      AICp       SBCp   PRESSp
1   2 1 0 0 0 0 0 0 0 0 0 0 74.99173 0.6012055 0.5985644 93.343234 -105.0982  -99.03732 76.91505
2   3 1 0 0 1 0 0 0 0 0 0 0 58.88904 0.6868371 0.6826616 43.305786 -140.0818 -130.99048 61.09347
3   4 0 1 1 1 0 0 0 0 0 0 0 52.54787 0.7205585 0.7149321 24.813669 -155.5131 -143.39139 56.02698
4   5 0 1 1 1 0 0 0 0 0 0 1 47.14481 0.7492911 0.7425152  9.353133 -170.1138 -154.96156 50.67951
5   6 1 1 1 1 0 0 0 0 0 0 1 45.33018 0.7589410 0.7507418  5.488987 -174.1191 -155.93650 49.32776
6   7 1 1 1 1 0 0 0 0 1 1 0 44.07020 0.7656414 0.7560102  3.417249 -176.4321 -155.21900 48.30509
7   8 1 1 1 1 1 0 0 0 1 1 0 43.87843 0.7666612 0.7553966  4.797520 -175.0993 -150.85579 48.92983
8   9 1 1 1 1 1 0 0 1 1 1 0 43.71541 0.7675281 0.7546130  6.270699 -173.6688 -146.39485 48.96623
9  10 1 1 1 1 1 0 0 1 1 1 1 43.65000 0.7678760 0.7532668  8.059312 -171.8979 -141.59353 49.35954
10 11 1 1 1 1 1 1 0 1 1 1 1 43.63379 0.7679622 0.7516215 10.006924 -169.9547 -136.61992 50.31343
11 12 1 1 1 1 1 1 1 1 1 1 1 43.63164 0.7679736 0.7498722 12.000000 -167.9623 -131.59700 50.60482
```

When running Best Subset Algorithm for a model with interaction terms where we look at the predictors using Happiness Score as the response variable, we see $Y \sim X_1 + X_2 + X_3 + X_4 + X_{27} + X_{37}$ seems to have the best values for C_p, AIC_p, R^2_{adj}, and $PRESS_p$ while a less complex model has a lower value of SBC_p. However, since SBC_p usually prefers more general models whereas AIC_p & C_p prefers more complex models and since the other criterions chose the same model as AIC_p & C_p, it seems that is the best model by the Best Subset Algorithm.

Based on this output from best subset algorithm, let's define our hypothesis.

$$H_O : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_{17} = \beta_{27} = \beta_{37} = \beta_{47} = 0$$

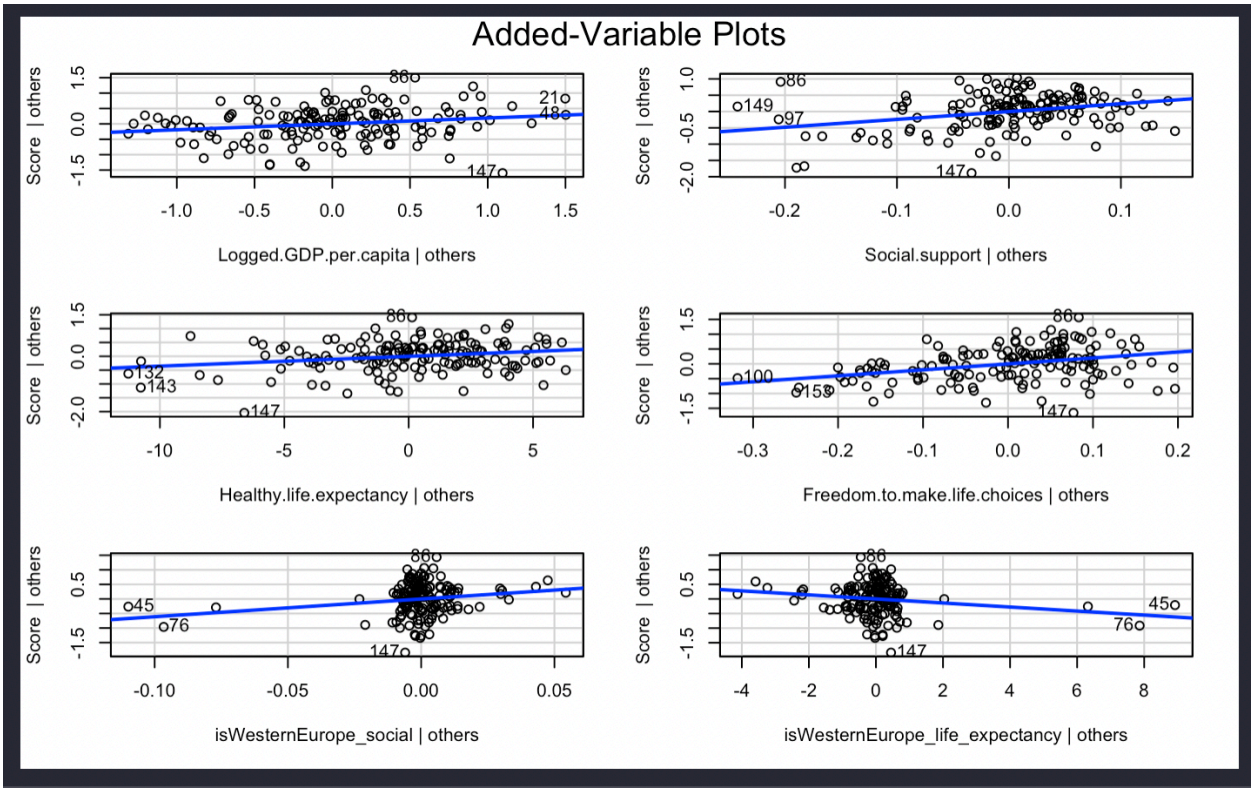
$$H_A : \text{at least one of } \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_{17}, \beta_{27}, \beta_{37}, \beta_{47} \neq 0$$

This is the model summary of the MLR model with interaction terms…

```
Call:
lm(formula = "Score ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy + Freedom.to.make.life.choices
+ isWesternEurope_social + isWesternEurope_life_expectancy",
    data = myDF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80824 -0.32851  0.05074  0.33587  1.40150

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.19453    0.47985  -4.573 1.02e-05 ***
Logged.GDP.per.capita           0.19107    0.07848   2.435 0.016117 *
Social.support                  2.40150    0.63173   3.801 0.000211 ***
Healthy.life.expectancy         0.03567    0.01285   2.776 0.006222 **
Freedom.to.make.life.choices    2.01278    0.44895   4.483 1.48e-05 ***
isWesternEurope_social          6.12036    2.59789   2.356 0.019807 *
isWesternEurope_life_expectancy -0.06931    0.03261  -2.126 0.035205 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5494 on 146 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:  0.756
F-statistic:  79.5 on 6 and 146 DF,  p-value: < 2.2e-16
```

Diagnostics:

AV Plot:



The AV plots indicate that a linear term of each of the predictors would be a helpful addition to the regression model including the interaction terms.

VIF Coefficients to judge Multicollinearity:

```
library(fmsb)
VIF(lm(Logged.GDP.per.capita ~ Social.support + Healthy.life.expectancy + Freedom.to.make.life.choices +
isWesternEurope_social + isWesternEurope_life_expectancy, myDF))
VIF(lm(Social.support ~ Logged.GDP.per.capita + Healthy.life.expectancy + Freedom.to.make.life.choices +
isWesternEurope_social + isWesternEurope_life_expectancy, myDF))
VIF(lm(Healthy.life.expectancy ~ Logged.GDP.per.capita + Social.support + Freedom.to.make.life.choices +
isWesternEurope_social + isWesternEurope_life_expectancy, myDF))
VIF(lm(Freedom.to.make.life.choices ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy +
isWesternEurope_social + isWesternEurope_life_expectancy, myDF))
VIF(lm(isWesternEurope_social ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy +
Freedom.to.make.life.choices + isWesternEurope_life_expectancy, myDF))
VIF(lm(isWesternEurope_life_expectancy ~ Logged.GDP.per.capita + Social.support + Healthy.life.expectancy +
Freedom.to.make.life.choices + isWesternEurope_social, myDF))
```

```
[1] 4.478409
[1] 2.964343
[1] 4.141441
[1] 1.408095
[1] 338.7635
[1] 338.8314
```

We can see that the VIF values for the interaction terms in the model are greater than 10 which indicates large multicollinearity from these terms. As such, we should remove them from the model. However, removing these terms from the model will give us a regular MLR model with

no interaction terms. As such, I will use the previous model found from Best Subset that was known to have satisfied all diagnostics without violation. Thus, my final model will be $Y \sim X_1 + X_2 + X_3 + X_4 + X_7$ which is the MLR model with no interaction terms.

I will validate this using k-fold cross validation below which was taken from the slides which resulted in these 3 folds.

	nvmax <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	5	0.558415	0.7604705	0.4441403	0.08349917	0.09457971	0.07384347

	nvmax <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	6	0.5599609	0.7649464	0.4452605	0.09135197	0.0679164	0.06328417

	nvmax <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	7	0.5583245	0.7639024	0.4424924	0.09313316	0.06004272	0.076035

All of the folds looked at had an R^2_{adj} of about 76% showing this model explains a lot of variance in Y.

Final Model Report and Application

The final model is $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7$. I thought that the linear relationships I discovered between many of the predictors in the scatter plot matrix during the exploratory analysis phase would cause multicollinearity issues in the MLR model unless a transformation or ridge regression was performed. However, there was no such issue with VIF coefficients being sufficiently below 10. I did find that there were some outliers based on the boxplots of certain predictors. This was confirmed by the leverage point analysis from the hat matrix and the influence plot. In the end, the number of outliers were small relative to the size of the dataset, I didn't remove any observations as there weren't too many outliers. In the future, a ridge regression model could be attempted to see if this reduces outliers.

1. *What predictor is the most important in predicting the mean response of Happiness score? Run diagnostics on the model to ensure no violations.*

$H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$
$F_S = \frac{MSR}{MSE} = \frac{113.054}{0.497} = 227.64 > F(1,151)$
As such, based on the result of the ANOVA F-test, we can say with 95% confidence that $\beta_1 \neq 0$.

2. *Which predictors in a multiple regression model without interaction terms can help predict the mean response of Happiness Score? Run relevant diagnostics and perform cross validation.*

$$H_O : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_A : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \neq 0$$

```
Analysis of Variance Table

Response: Score
                            Df  Sum Sq Mean Sq F value    Pr(>F)
Logged.GDP.per.capita        1 113.054 113.054 363.513 < 2.2e-16 ***
Social.support               1  12.198  12.198  39.221 3.946e-09 ***
Healthy.life.expectancy      1   4.561   4.561  14.665 0.0001896 ***
Freedom.to.make.life.choices 1   8.571   8.571  27.558 5.240e-07 ***
isWesternEurope              1   3.944   3.944  12.683 0.0004977 ***
Residuals                  147  45.718   0.311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 457.6463
[1] TRUE
```

$$F_S = \frac{MSR}{MSE} = \frac{142.328}{0.311} = 457.6453 > F(p-1, n-p)$$

As such, based on the result of the ANOVA F-test, we can say with 95% confidence that at least one of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \neq 0$

On the well reputed Data Science publication, Towards Data Science, Lina Faik also used regression to model happiness as it allowed her to quantify the relationships between some variables and also its relevance relationships through different statistical tests. It also allowed for controlling certain variables during analysis in order to understand the unique effect of certain variables. As such, it seems as though using regression techniques to model Happiness works very well to answer these research questions.

I believe the most important part of this research project was determining the variables that help increase happiness. Increasing happiness in countries can improve the lives of many citizens and reduce deaths of despair which is on the rise due to coronavirus. Showing that making your citizens wealthier and investing in their health creates happiness is extremely important so that policymakers can use this information to develop policies to improve happiness/well-being of all.

References

7 reasons why happiness is important in your life. (2021, January 05). Retrieved April 29, 2021, from https://www.happierhuman.com/why-happiness-important/

Durham, J. (2020, June 17). Why is happiness so important? Retrieved April 29, 2021, from http://www.lifecoachexpert.co.uk/whyishappinesssoimportant.html

Faik, L. (2020, August 23). Understanding happiness dynamics with machine learning (part 2). Retrieved April 29, 2021, from https://towardsdatascience.com/understanding-happiness-dynamics-with-machine-learning-part-2-4df36e52486

Singh, A. (2021, March 22). World happiness REPORT 2021. Retrieved April 29, 2021, from https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021