Group 16: World Happiness Report Analysis
Shiun Choi, Andy Ho, Evelyn Isaka, Chloe Kim, Kyoka Ono, Jialing Yuan

INTRODUCTION

In this study, various measures of well-being were utilized to study happiness across many nations and to create a predictive model to analyze how the different measures affect the overall happiness score.

Experts across numerous fields believe indicators of happiness and well-being can help analyze the progress of countries. The statistics in this report come from the World Happiness Report data which uses scores and rankings from the Gallup World poll. These scores are based on answers from one main question regarding life evaluation. The report reviews the state of happiness and how science explains different variations in happiness. The data was released by the United Nations which ranks countries in the order of their happiness levels. Additionally, the report is gaining global recognition as people increasingly begin to use happiness indicators to influence policies.

We will look at seven total variables, six predictor variables and one response variable. The six predictor variables will be GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. We will observe how these six predictors affect the one response variable, which is the overall score. The data includes 156 observations since there were 156 countries studied.

In order to create the model, we first transformed all of the variables from the dataset using the Box-Cox transformation so that they would follow normal distributions. We then used the results from subset regression as well as forwards and backwards regression to determine which variables would be present in the final model. Finally, we verified that the final model was appropriate by examining the diagnostic plots.

We will first look at the individual variables and their relationships with each other. Then, we will explain the initial model we started with, and how we changed this model in order to create the final model. Finally, we will assess the model using diagnostic tools, and discuss the real-world context behind the dataset and model.

DATA DESCRIPTION

Figures 1.1 and 1.2 show the summary statistics (IQR, mean) and standard deviation of the data, respectively. To draw an overall pattern and compare these summary statistics across variables, we created a boxplot, as shown in Figure 1.3. Figure 1.4 shows the density plot of variables, from which we observe that most are normal distributions, but some are skewed to the left or right.

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

Figure 1.1: IQR and mean of the data

```
                          Score                GDP.per.capita
                     1.11311987                    0.39838946
                 Social.support      Healthy.life.expectancy
                     0.29919140                    0.24212400
Freedom.to.make.life.choices                    Generosity
                     0.14328947                    0.09525444
    Perceptions.of.corruption
                     0.09453784
```

Figure 1.2: Standard deviation of the data

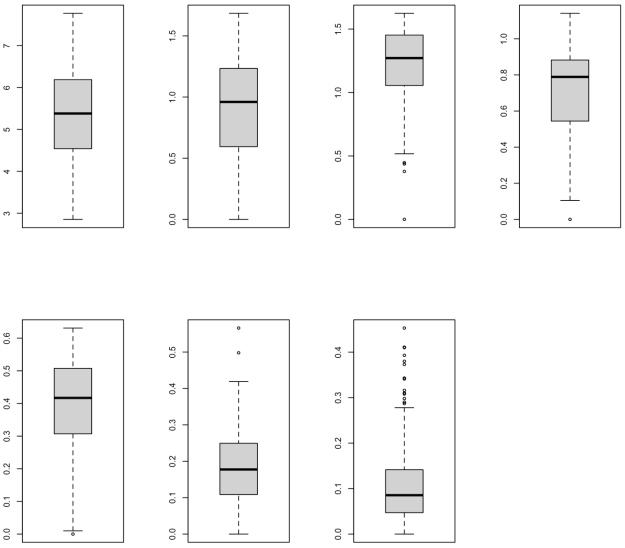

Figure 1.3: Boxplot of variables

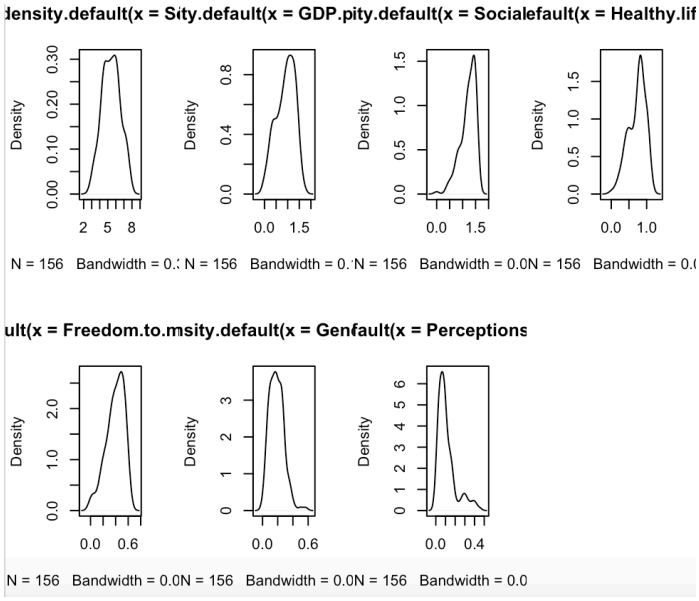

Figure 1.4: Density plot of variables

From the scatterplot matrix in Figure 1.5, we can see that most of the predictors form a positive linear relationship with the response variable and some with each other. Generosity forms a weaker relationship with Score compared to the other variables. Also, we see in Figure 1.6 that the correlation matrix exhibits pretty high correlation between GDP.per.capita and Social.support, GDP.per.capita and Healthy.life.expectancy, and Social.support and Healthy.life.expectancy. Because of this, we should consider checking for multicollinearity when building our model.

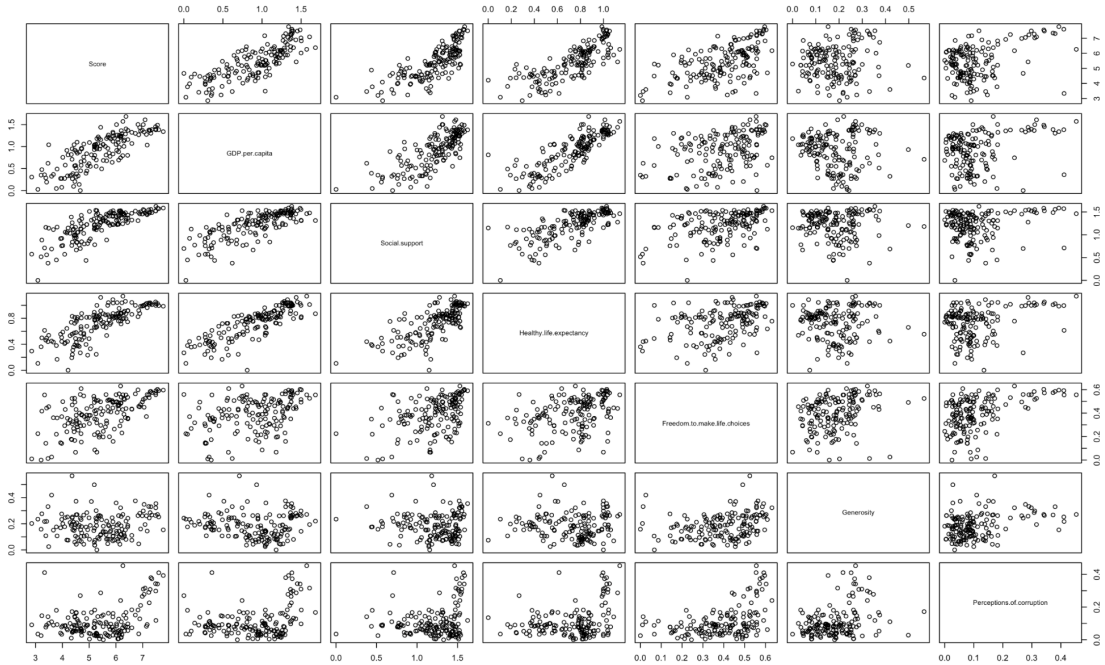

Figure 1.5: Scatterplot matrix of all variables

	Score	GDP.per.capita	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Score	1.00000000	0.79388287	0.77705779	0.77988315	0.5667418	0.07582369	0.3856131
GDP.per.capita	0.79388287	1.00000000	0.75490573	0.83546212	0.3790791	-0.07966231	0.2989198
Social.support	0.77705779	0.75490573	1.00000000	0.71900946	0.4473332	-0.04812645	0.1818995
Healthy.life.expectancy	0.77988315	0.83546212	0.71900946	1.00000000	0.3903948	-0.02951086	0.2952828
Freedom.to.make.life.choices	0.56674183	0.37907907	0.44733316	0.39039478	1.0000000	0.26974181	0.4388433
Generosity	0.07582369	-0.07966231	-0.04812645	-0.02951086	0.2697418	1.00000000	0.3265375
Perceptions.of.corruption	0.38561307	0.29891985	0.18189946	0.29528281	0.4388433	0.32653754	1.0000000

Figure 1.6: Correlation matrix of all variables

RESULTS AND INTERPRETATION

INITIAL MODEL

Since all variables are continuous, we start with multiple linear regression. Our initial variable includes all predictor variables (GDP.per.capita, Social.support, Healthy.life.expectancy, Freedom.to.make.life.choices, Generosity, Perceptions.of.corruption) and the response variable (Score). From the summary table (Figure 2.1), we see a relatively good R-squared and a statistically significant F-statistic with p-value less than 0.05. However, we notice that two predictors are not statistically significant, which are Generosity and Perceptions.of.corruption. Then, looking at diagnostic models (Figure 2.2), we especially observe a large curvature in the Residuals vs Fitted graph, and several heavy tails in the Normal Q-Q plot. These construct a slight violation of model assumptions, and thus we take advanced steps to improve our model.

VIF is generally good since no variable has VIF exceeding 5 (Figure 2.3). However, since there are 6 predictors and our data description showed that there was high correlation between some of the

predictor variables, we will try model selection and seek for a possible better model to explore the core factors of happiness and decrease multicollinearity.

Candidate 1: We attempted to transform the response variable and the predictor variables simultaneously, by means of Box-Cox method. The R-output suggested for transformation in all the variables except for the Score and GDP.per.capita variables (Figure 2.4). As a result, Adjusted R-squared improved by around 2.5%; however, Generosity and Perception.of.corruption still remained insignificant, as shown in Figure 2.5.

Candidate 2: We transformed the response variable and the predictor variables separately, by using Inverse Response Plot for the former and the Box-Cox method for the latter. As a result, the R-output suggested transformation for every variable except for GDP.per.capita (Figure 2.6). Additionally, the Inverse Response Plot suggested using ^0.84 since it has the least RSS (Figure 2.7). The transformation showed no improvement in Adjusted R-squared. In fact, it decreased slightly compared to the transformed model 1. The model is very similar to the first candidate, in which the significance of GDP.per.capita increased; however, Generosity and Perception.of.corruption still remained insignificant, as shown in Figure 2.8.

Candidate 3: Lastly, we conducted forward and backward stepwise regressions with AIC and BIC, as well as considered all the subset models. Both Backward AIC and BIC, and Forward BIC suggested a model with four predictors: GDP per capita, social support, healthy life expectancy, and freedom. Additionally, after considering all the possible subset models, the values for AIC, AICc, BIC were the lowest for a model with four predictors (GDP per capita, social support, healthy life expectancy, and freedom to make life choices) whereas Adjusted R-squared suggested the best model to be the model with five predictors (GDP per capita, social, healthy life expectancy, freedom, and generosity) (Figure 2.9).

FINAL MODEL

To select our final model, we performed three methods of variable selection: forwards selection (Figure 2.9) and backwards selection (Figure 2.9) with AIC and BIC, and the subset selection with R^2, AIC, and BIC. The stepwise selection gave us the same results, and the subset selection method also showed similar results. We also conducted an ANOVA test comparing the reduced model (which excludes generosity and perception) and full model with all 6 predictors. The partial F test had a p value of 0.3096 (>0.05) and therefore we fail to reject the null hypothesis and keep the reduced model.

Therefore, our final model is:

$$y^{0.8} = 0.3299*GDP.per.capita + 0.345*social.support^2 + 0.6027*healthy.life.expectancy^{1.6} + 1.1724*freedom^{1.6} + e$$

DIAGNOSTIC TOOLS

The statistical summary of our final model with the 4 variables GDP per capita, social support, healthy life expectancy, and freedom to make life choices was the best predictive model for the happiness score. This model has an R^2 of 0.7907, R^2adj of 0.7852, AIC of 375.15, and F-statistic of 142.6 with a p-value smaller than 2.2e-16, indicating a good performance. Furthermore, when we look at diagnostic models, almost all model assumptions of multi-linear regression are satisfied. In Figure 2.10, we see a relatively flat and straight line in both Residuals vs Fitted plot and square root of standardized residual plot, implying a good linearity and constant variance. In the Normal Q-Q plot, it shows the normality of errors. Since $2*(p+1)/n = 0.178$ since we have p=4 and n=156, we observe no bad leverage points. In addition, in the standardized residual plots (Figure 2.11), points are scattered randomly around the horizontal axis and variability is also constant, showing that the model is valid. Finally, in the added-variable plots (Figure 2.12), we see a clear slope of lines in four graphs, which indicates they are all significant when contributing to the response variable.

DISCUSSION

SUMMARY OF PROJECTS

Due to the interest in figuring out the main factors that affect people's happiness, we found data posted by the United Nations in 2019. The dataset includes 156 observations since it collects people's ratings of their home countries, with a total of 156 countries. There are 7 variables, including overall happiness rating as the response variable, and other 6 predictors that serve as measurements to predict Y. We first observed the summary statistics of the variables. To get a clear understanding, we visualized them by using box plots and density plots to observe their means, standard deviations, and overall distribution patterns. It turns out that all variables form a normal distribution, but some have a right or left skewness. Since all variables are continuous and the correlation matrix shows a clear positive relationship between variables, we decided to use multi-linear regression as the basic model. While the basic model is good in R squared, there exist certain violations against the model assumption. Therefore, we tried to use transformation to seek a better model.

Simultaneously applying the Box-Cox transformation to both variables yielded a slight increase in the R-squared value and an enhancement in diagnostic model performance, particularly evident in the residual plot. Subsequently, we formulated an alternative model where predictors were transformed first, followed by the utilization of inverse response plots to address the response variable. Although this model exhibited results similar to the initial model, it demonstrated a slightly more linear trend in the residual plot. Thus, we proceeded with model 2 for further analysis.

Despite having a VIF where each variable was less than 5, we still did model selection to both reduce complexity and figure out key factors that influence happiness. After conducting both stepwise and subset model selection, we confirmed by the ANOVA test that the model with 4 variables is the most appropriate, and its R squared and diagnostic model are both valid. Therefore, we finish our analysis for this dataset.

REAL-LIFE SITUATION

By intuition, the four variables: GDP per capita, social support, freedom to make life choices, and healthy life expectancy are indeed essential factors that influence people's feelings toward happiness. GDP directly relates to salary and price of commodities; social support decides how warm-hearted the overall social environment is, which affects mental health; freedom to make life choices gives people the confidence to realize their dreams and willings during any time of life; healthy life expectancy is most important since it decides life quality and aspiration toward future. Thus, the higher scores these 4 factors are, the happier the people of certain countries will be. On the other hand, the perception of corruption and generosity are also meaningful, but they do not directly influence common people's lives and thus do not serve as significant as the other four factors.

There are also many articles that discuss how factors such as GDP and health insurance system increase people's life satisfaction. For example, in an article published by CEPR, which is a London-based European network of economists, the writer points out that "people in countries with a GDP per capita of below $6,700 were 12% less likely to report the highest level of life satisfaction than those in countries with a GDP per capita of around $20,000", clearly indicating a positive correlation between life satisfaction and GDP. In addition, research named "Subject Well-being, Income, Economic Development, and Growth" published by BROOKINGS conducted precise analysis from datasets across countries. This research applied both statistical and economic methods to discover the relationship

between GDP, income, and life satisfaction. One of the results claimed “We find that happiness is positively related to per capita GDP across a sample of 69 countries”, and showed that some additional measures, such as enjoyment and love, are all higher in countries with higher GDP per capita. Regarding health, a study in Norway before and during the pandemic shows the result that there exists a negative relationship between ill health and life satisfaction. The study claims that a person with poor health is more likely to experience a worsened working situation, which leads to lower life satisfaction. Similarly, there are many studies indicating the impact of social support and freedom on life satisfaction. A passage on NIH claims that among stressed and somatized communities, social support, especially that from family, is significant for helping them become optimistic. To conclude, these four factors do show a large influence on life satisfaction, which leads to happiness.

WEAKNESSES

However, there are still several weaknesses that need further improvement. First of all, this data is from 2019, when there was no COVID-19 pandemic. Therefore, the overall life quality and perception of happiness may be high due to peaceful times. However, during and after the pandemic, social and economic problems arose, potentially leading to a lower score on the same survey. As a result, the conclusion from this dataset cannot be generalized to later years such as 2020 to 2022. Secondly, since we did not apply log transformation, it’s hard to make a reasonable interpretation according to coefficients. In addition, the survey only includes 6 variables, and there exists a probability that some other factors, such as education, also have a significant impact on the perception of happiness but were not included.

For further exploration, after we learn more models, we can hopefully apply a better one that both fits the data and clearly interprets coefficients to make results more meaningful. Also, it may be a good idea to compare our conclusion with that of group 8, which did an analysis of the dataset from 2020. The comparison may show certain influences of how the pandemic changes people’s perception of happiness. For example, the coefficient of freedom in our final model is the largest, but possibly after the pandemic, people noticed the significance of health and thus rated health factors to be more important.

Overall, we hope everyone is happy every day!

Appendix

Graphs:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.75304 -0.35306  0.05703  0.36695  1.19059

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.7952     0.2111   8.505 1.77e-14 ***
GDP.per.capita                   0.7754     0.2182   3.553 0.000510 ***
Social.support                   1.1242     0.2369   4.745 4.83e-06 ***
Healthy.life.expectancy          1.0781     0.3345   3.223 0.001560 **
Freedom.to.make.life.choices     1.4548     0.3753   3.876 0.000159 ***
Generosity                       0.4898     0.4977   0.984 0.326709
Perceptions.of.corruption        0.9723     0.5424   1.793 0.075053 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 149 degrees of freedom
Multiple R-squared:  0.7792,    Adjusted R-squared:  0.7703
F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

Figure 2.1: Summary table of initial model

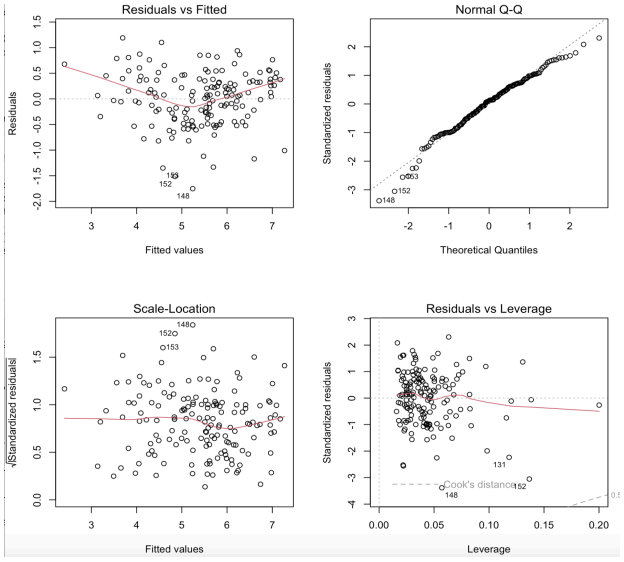

Figure 2.2: Diagnostic plots of initial model

```
                GDP.per.capita                  Social.support      Healthy.life.expectancy
                      4.115838                       2.735651                       3.572728
Freedom.to.make.life.choices                     Generosity      Perceptions.of.corruption
                      1.575090                       1.224101                       1.431594
```

Figure 2.3: VIF results for initial model

```
bcPower Transformations to Multinormality
                           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Score                         1.2155        1.00       0.6717       1.7592
GDP.per.capita                1.1675        1.00       0.9401       1.3949
Social.support                2.5554        2.00       1.9832       3.1276
Healthy.life.expectancy       1.6646        2.00       1.3148       2.0144
Freedom.to.make.life.choices  1.6095        1.61       1.2563       1.9627
Generosity                    0.4388        0.50       0.2004       0.6772
Perceptions.of.corruption     0.2966        0.33       0.1511       0.4421
```

Figure 2.4: Suggested power for transformed model 1

```
Call:
lm(formula = Score ~ GDP.per.capita + social + healthy + freedom +
    generosity + perception)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5088 -0.3016  0.0099  0.3264  1.2570

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.6024     0.2183  11.920  < 2e-16 ***
GDP.per.capita   0.5680     0.2182   2.603 0.010162 *
social           0.6172     0.1096   5.632 8.63e-08 ***
healthy          1.0002     0.2497   4.006 9.72e-05 ***
freedom          1.7227     0.4434   3.885 0.000153 ***
generosity       0.4978     0.3944   1.262 0.208867
perception       0.2930     0.3669   0.799 0.425725
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5106 on 149 degrees of freedom
Multiple R-squared:  0.7977,    Adjusted R-squared:  0.7896
F-statistic: 97.94 on 6 and 149 DF,  p-value: < 2.2e-16
```

Figure 2.5: Summary table of transformed model 1

-

```
bcPower Transformations to Multinormality
                           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
GDP.per.capita                1.1304        1.00       0.9027       1.3581
Social.support                2.4078        2.00       1.8287       2.9870
Healthy.life.expectancy       1.5832        1.58       1.2300       1.9364
Freedom.to.make.life.choices  1.6408        1.64       1.2837       1.9979
Generosity                    0.4582        0.50       0.2208       0.6956
Perceptions.of.corruption     0.2975        0.33       0.1518       0.4432

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                    LRT df       pval
LR test, lambda = (0 0 0 0 0 0) 494.6002  6 < 2.22e-16

Likelihood ratio test that no transformations are needed
                                    LRT df       pval
LR test, lambda = (1 1 1 1 1 1) 151.3441  6 < 2.22e-16
```

Figure 2.6: Suggested power for transformed model 2

```
      lambda      RSS
1  0.8477321 33.00193
2 -1.0000000 40.05805
3  0.0000000 34.44703
4  1.0000000 33.04567
```

Figure 2.7: Inverse Response plot of Y

```
Call:
lm(formula = score2 ~ GDP.per.capita + social + healthy2 + freedom +
    generosity + perception)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91407 -0.16684  0.00977  0.19033  0.72376

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.1917     0.1240  17.682  < 2e-16 ***
GDP.per.capita   0.3455     0.1263   2.737 0.006964 **
social           0.3560     0.0635   5.606 9.74e-08 ***
healthy2         0.5872     0.1571   3.737 0.000264 ***
freedom          0.9858     0.2568   3.839 0.000182 ***
generosity       0.2720     0.2281   1.192 0.234980
perception       0.1545     0.2114   0.731 0.465977
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2954 on 149 degrees of freedom
Multiple R-squared:  0.794,     Adjusted R-squared:  0.7857
F-statistic:  95.7 on 6 and 149 DF,  p-value: < 2.2e-16
```

Figure 2.8: Summary table of transformed model 2

```
Step:  AIC=-375.15
score2 ~ GDP.per.capita + social + healthy2 + freedom

                Df Sum of Sq    RSS     AIC
<none>                      13.210 -375.15
- GDP.per.capita  1   0.61982 13.830 -370.00
- healthy2        1   1.29186 14.502 -362.59
- freedom         1   2.40025 15.610 -351.10
- social          1   2.69131 15.901 -348.22
Step:  AIC=-375.15
score2 ~ social + healthy2 + freedom + GDP.per.capita

              Df Sum of Sq    RSS     AIC
<none>                    13.210 -375.15
+ generosity  1  0.159627 13.050 -375.04
+ perception  1  0.082156 13.128 -374.12
```

```
Step:  AIC=-359.9
score2 ~ GDP.per.capita + social + healthy2 + freedom

                Df Sum of Sq    RSS     AIC
<none>                      13.210 -359.90
- GDP.per.capita  1   0.61982 13.830 -357.80
- healthy2        1   1.29186 14.502 -350.39
- freedom         1   2.40025 15.610 -338.90
- social          1   2.69131 15.901 -336.02
```

	Size	Radj2	AIC	AICc	BIC
1	1	0.6505085	-302.1688	-302.0130	-296.0691
2	2	0.7402613	-347.4865	-347.2250	-338.3369
3	3	0.7765636	-369.9951	-369.6004	-357.7957
4	4	0.7851644	-375.1483	-374.5920	-359.8991
5	5	0.7863455	-375.0449	-374.2983	-356.7458
6	6	0.7856800	-373.6033	-372.6368	-352.2543

Figure 2.9 Stepwise and model selection results

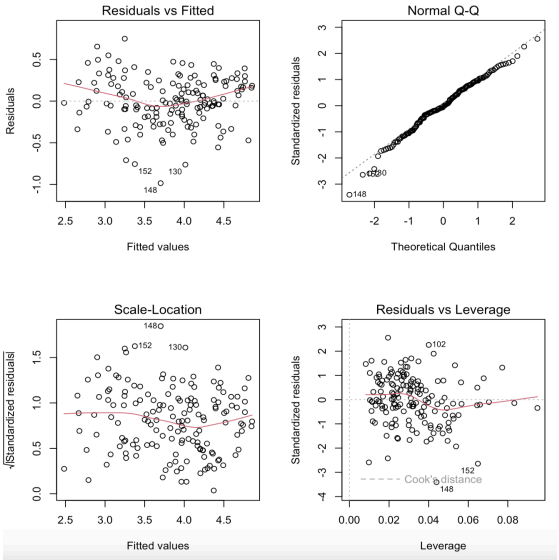

Figure 2.10: Diagnostic plots of final model

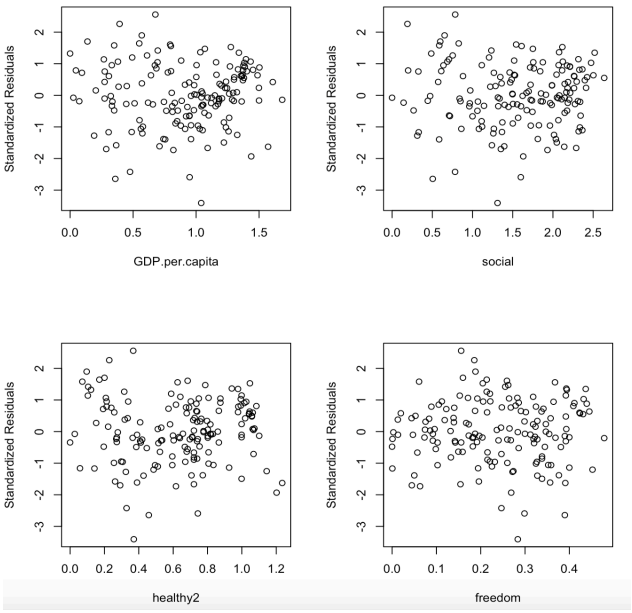

Figure 2.11: Standardized residual plots of final model

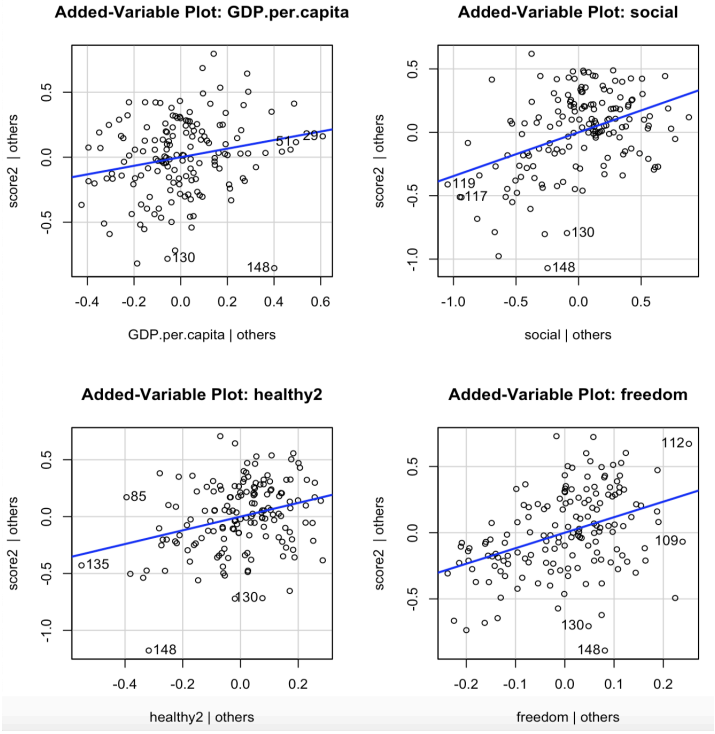

Figure 2.12: Added variable plots of final model

Data Resource:
https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv

Reference:
1. Rustichini, Also, and Eugenio Proto. "GDP and Life Satisfaction: New Evidence." CEPR, 11 Jan. 2014, cepr.org/voxeu/columns/gdp-and-life-satisfaction-new-evidence
2. Bakkeli NZ. Health, work, and contributing factors on life satisfaction: A study in Norway before and during the COVID-19 pandemic. SSM Popul Health. 2021 May 4;14:100804. doi: 10.1016/j.ssmph.2021.100804. PMID: 34027009; PMCID: PMC8129931.
3. Ali A, Deuri SP, Deuri SK, Jahan M, Singh AR, Verma AN. Perceived social support and life satisfaction in persons with somatization disorder. Ind Psychiatry J. 2010 Jul;19(2):115-8. doi: 10.4103/0972-6748.90342. PMID: 22174534; PMCID: PMC3237127.
4. Sacks, Daniel, et al. "Subjective Well-Being, Income, Economic Development and Growth." Brookings, 1 Oct. 2010, www.brookings.edu/articles/subjective-well-being-income-economic-development-and-growth/.