Happiness

Writer: Qintao Jia, Anna Hyunjung Kim

I. Introduction

Background:

The data focuses on understanding the factors that influence national happiness levels across different countries. This topic is important because happiness is increasingly recognized as a key indicator of social progress, complementing traditional economic measures such as GDP.

Data Collection:

This study investigates the factors that explain differences in happiness levels across countries. The primary data source is the World Happiness Report(https://worldhappiness.report/). It is based on responses collected through the Gallup World Poll. Each year, Gallup surveys approximately 1,000 adults in more than 160 countries using a consistent methodology to ensure comparability across countries and over time.

Other important factors, like GDP per capita, healthy life expectancy, social support, freedom, generosity, and corruption perception, were also measured. These came from trusted sources such as the World Bank, the World Health Organization, and additional Gallup surveys.

When recent data is missing, the researchers used older data or made careful estimates based on related information. However, these estimated numbers were only used for analysis; not for the official happiness rankings. The rankings are based only on real collected data.

DATA SET INFORMTAION

Data source: Sazidthe1. (2023). World Happiness Report (till 2023) [Data set]. Kaggle. https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors

Data Originally from https://worldhappiness.report/

1. **The World Happiness Report (2020)**

    a. **153 unique observations**

    b. **9 columns**

        i. **country**
        ii. **region**
        iii. **happiness_score**
        iv. **gdp_per_capita**
        v. **social_support**
        vi. **healthy_life_expectancy**
        vii. **freedom_to_make_life_choice**
        viii. **generosity**
        ix. **perceptions_of_corruption**

2. **The World Happiness Report (2015)**

    a. **158 unique observations**

    b. **9 columns**

        i. **country**
        ii. **region**
        iii. **happiness_score**
        iv. **gdp_per_capita**
        v. **social_support**
        vi. **healthy_life_expectancy**
        vii. **freedom_to_make_life_choice**
        viii. **generosity**
        ix. **perceptions_of_corruption**

Variable	**Description**	**Type**	**Unit**
country	Name of the country	Categorical	N/A
region	Region the country belongs to	Categorical	N/A
happiness_score	Average life evaluation score based on the Cantril ladder (0–10 scale)	Quantitative	Score (0–10)
gdp_per_capita	Log of GDP per capita based on Purchasing Power Parity (PPP) in constant 2021 international dollars	Quantitative	Log(USD, PPP-adjusted)
social_support	National average of the binary response (0 = no, 1 = yes) to the question: “Do you have someone to rely on in times of trouble?”	Quantitative (proportion)	Proportion (0–1)
healthy_life_expectancy	Life expectancy at birth adjusted for health quality	Quantitative	Years
freedom_to_make_life_choices	National average of responses to: “Are you satisfied with your freedom to choose what you do with your life?” (0 = no, 1 = yes)	Quantitative (proportion)	Proportion (0–1)
generosity	Residual from regressing donation behavior on GDP per capita; measures prosocial behavior	Quantitative	Index (no unit)
perceptions_of_corruption	National average of responses to corruption-related questions (0 = not widespread, 1 = widespread)	Quantitative (proportion)	Proportion (0–1)

Research Questions:

a. Do the happiness scores of people on different continents in 2020 differ significantly (Q1)

b. Have the happiness scores of people in various countries changed between 2015 and 2020 (before and after the COVID-19) (Q2)

c. How do the different explanatory variables influence the happiness score in 2020 (Q3)

TENTATIVE Plan of Action:

a. We will use descriptive statistics and boxplots to explore the data. These summaries will help us understand the distribution, variability, and outliers in the data.

b. We will use the Shapiro-Wilk test and Q-Q plots to assess whether the variables are normally distributed (Normality). Moreover, we will use Levene's test for ANOVA to verify homogeneity of variances. Lastly, Residual vs. fitted plots and Q-Q plots of residuals will be used for linearity and normality of residuals.

c. Statistical procedures.

Q1: ANOVA

Q2: Pair t-test

Q3: Multiple regression

II. Statistical Procedures Used

Question 1

Descriptive data (Figure 1.1) and Box Plot (Figure 1.2) were made to get the summary of the happiness score among different countries in 2020. Histogram (Figure 1.3) was prepared to visualize the distribution of data. QQ plot (Figure 1.4) and Shapiro-Wilk normality test (Figure 1.5) were used to assess whether the data comes from a normally distributed population. Levene's Test (Figure 1.6) was conducted to check the assumption of homoscedasticity. ANOVA test (Figure 1.7) was applied to compare the mean difference.

Question 2

We used boxplots (Figure 2.1) to compare the distribution of happiness scores between the years 2015 and 2020. The plots suggest the median of happiness scores slight increase in 2020 compared to 2015. But the minimum decreased. It's not clear so we should use a Pair t-test. QQ plots (Figure 2.2) were also used to assess normality for happiness scores in both years individually, and for the differences between paired scores. The points are followed the line. It means the normality assumption is fine. Histograms (Figure 2.3) were used for each year to see the distribution shape. It appeared roughly symmetric in both years, though slightly left-skewed in 2020. Also, the normality of the differences between 2015 and 2020 happiness scores was tested using a Shapiro-Wilk test (p-value = 0.1116). We used a pair of t-test.

Question 3

Residual plots (Figure 3.1) were made after fitting multiple regression models to check the assumptions of linearity and homoscedasticity. The residuals vs. fitted plot showed no clear non-linear patterns, so it suggests the linearity assumption. Q-Q plots (Figure 3.1) of the residuals were also examined to assess normality. It indicated that the residuals were approximately normally distributed. To see the multicollinearity, we calculated the Variance Inflation Factors (VIF) (Figure 3.2)for each predictor in Model 60. All VIF values were below 5. That means we could say there isn't multicollinearity. We used multiple regression.

III. Summary of Statistical Findings

Question 1

The median happiness score increases with the order from sub-Sahara Africa, south Asia to West Europe and North America. The QQ-plot and p-value =0.1635 in Shapiro-Wilk normality test clearly show the data are normalized distribution. The Levene's test shows equal variances across different regions. By using the ANOVA test, the p-value is insignificant when comparing the happiness score of North America and other regions. The only one exception is between North America and West Europe where the p-value is 0.48, which means we can moderately assert that the happiness score are same between North America and West Europe.

Question 2

A paired t-test was computed to compare happiness scores for the 149 countries that had data available in both 2015 and 2020.The test is statistically significant (t = -2.34, p = 0.021) It indicates that the mean happiness score in 2020 was significantly lower than in 2015. The mean difference is -0.112. Also, the 95% confidence interval for the difference doesn't have 0 (from -0.207 to -0.017).

For curiosity, we also conducted paired t-tests within each region to compare happiness scores between 2015 and 2020. However, because of the small sample sizes in each region, the results were not statistically meaningful. Our paired t-test was only significant when performed on the full dataset having all countries.

Question 3

To build a multiple regression model, we began by fitting simple linear regressions (SLR) with each variables, then tested combinations of two up to six variables (Figure 3.3). Starting from models with five variables, the results became statistically strong. And there was not much improvement in how well the model explained happiness when we added a sixth variable. Among the variables, **generosity** appeared to be statistically insignificant in most models. so, we determined that the most appropriate model was the five-variable combination excluding **generosity (model 51)(Figure 3.4) .** To confirm the best combination of variables, we applied the step function using the step() function with AIC. The step selected the same five-variable model (model 51).

The model has adjusted R-squared as 0.737. It’s a powerful number. All five variables were statistically significant. Also, the F-statistic shows that the overall model was highly significant (F = 86.25, $p < 0.0001$).

IV. Scope of Inference

This dataset allows cross-national comparisons of subjective happiness and related factors.

However:

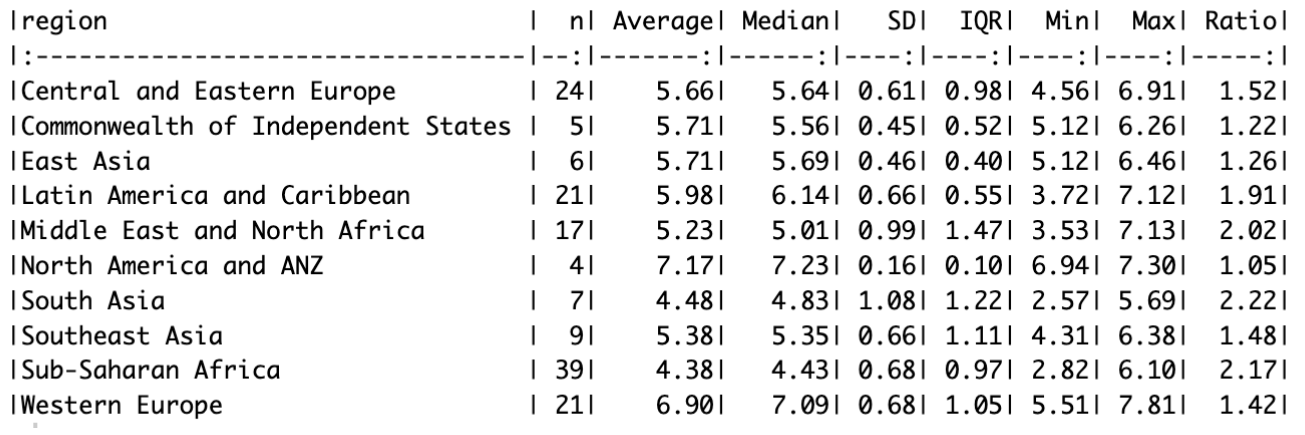
(a) It reflects correlational relationships not causal ones. Countries were not randomly assigned to different levels of GDP, social support, or other variables. As a result, we cannot conclude that, for example, increasing GDP causes an increase in happiness.

(b) It cannot be generalized beyond the countries included in the dataset. Although Gallup surveys are nationally representative within each country, the countries themselves were not randomly selected from the world population.

These limitations indicate that while our findings highlight important patterns, they do not provide definitive evidence of causality. Further experimental research would be needed to make causal structure.

V. Appendix

Figure 1.1. Q1 Summary



```
|region                            |  n| Average| Median|   SD|  IQR|  Min|  Max| Ratio|
|:---------------------------------|--:|-------:|------:|----:|----:|----:|----:|-----:|
|Central and Eastern Europe        | 24|    5.66|   5.64| 0.61| 0.98| 4.56| 6.91|  1.52|
|Commonwealth of Independent States |  5|    5.71|   5.56| 0.45| 0.52| 5.12| 6.26|  1.22|
|East Asia                         |  6|    5.71|   5.69| 0.46| 0.40| 5.12| 6.46|  1.26|
|Latin America and Caribbean       | 21|    5.98|   6.14| 0.66| 0.55| 3.72| 7.12|  1.91|
|Middle East and North Africa      | 17|    5.23|   5.01| 0.99| 1.47| 3.53| 7.13|  2.02|
|North America and ANZ             |  4|    7.17|   7.23| 0.16| 0.10| 6.94| 7.30|  1.05|
|South Asia                        |  7|    4.48|   4.83| 1.08| 1.22| 2.57| 5.69|  2.22|
|Southeast Asia                    |  9|    5.38|   5.35| 0.66| 1.11| 4.31| 6.38|  1.48|
|Sub-Saharan Africa                | 39|    4.38|   4.43| 0.68| 0.97| 2.82| 6.10|  2.17|
|Western Europe                    | 21|    6.90|   7.09| 0.68| 1.05| 5.51| 7.81|  1.42|
```

Figure 1.2. Q1 Box Plot

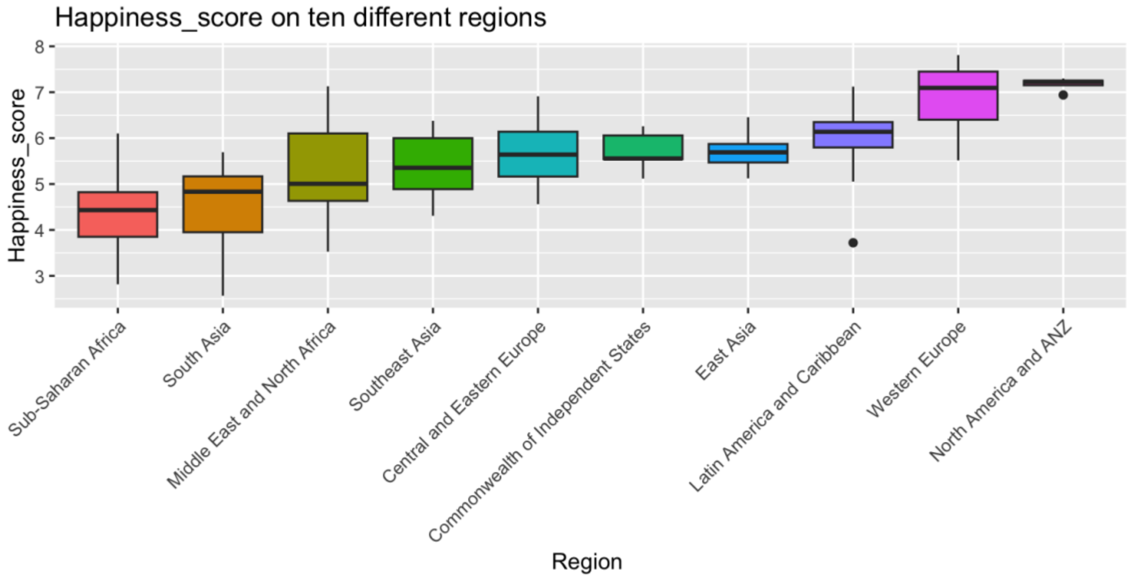


Figure 1.3. Q1 Histogram

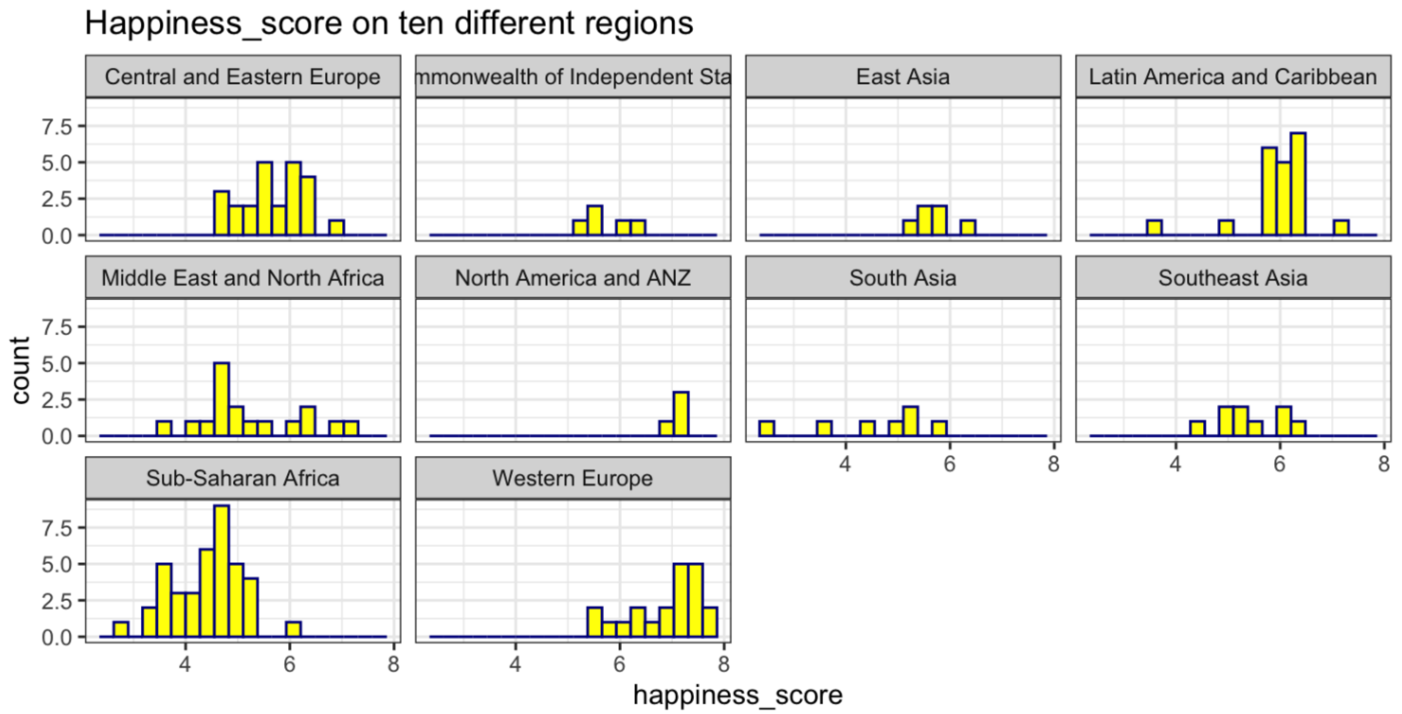


Figure 1.4. Q1 QQ Plot

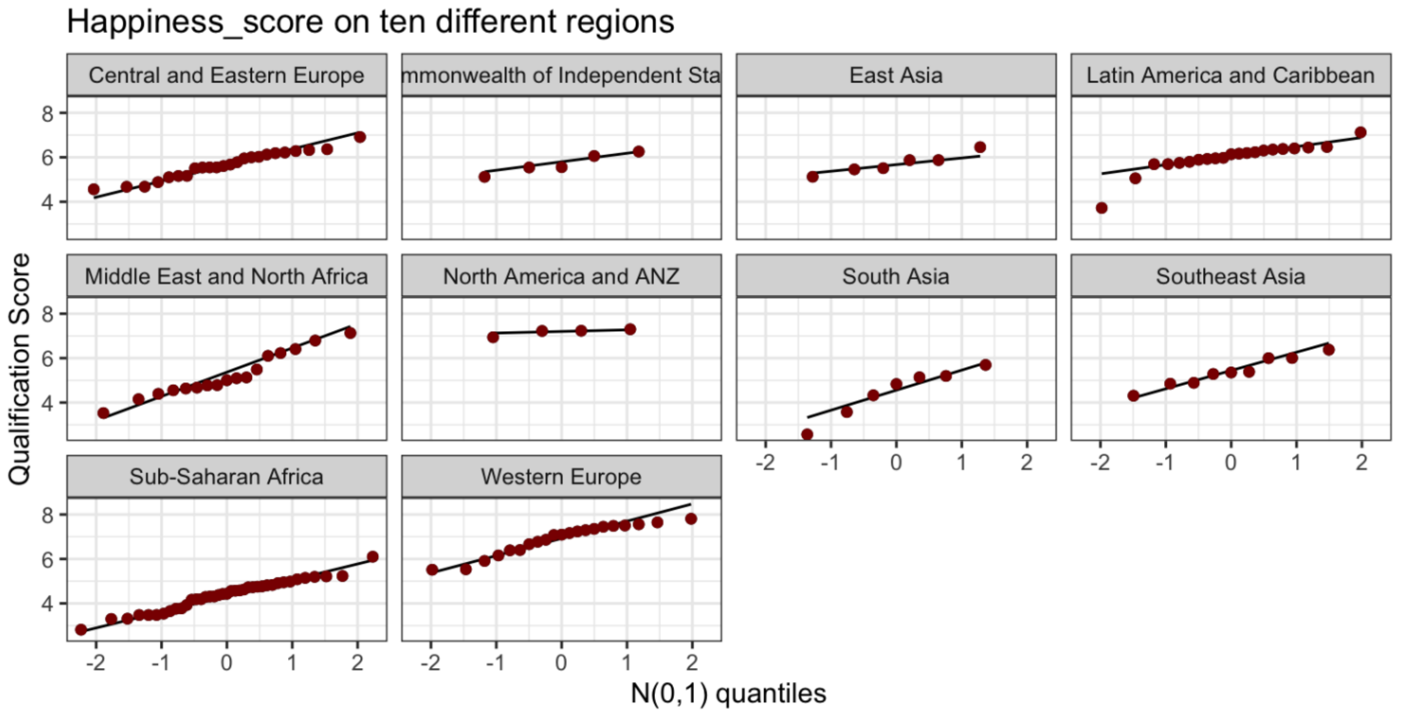


Figure 1.5. Q1 Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  model.fit$residuals
W = 0.986986, p-value = 0.1635
```

Figure 1.6 Q1 Levene's Test

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   9 1.45152 0.17173
      143
```

Figure 1.6 Q1 ANOVA

```
  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$region
                                                                        diff       lwr.ci       upr.ci
Commonwealth of Independent States-Central and Eastern Europe    0.048309958 -0.643703847  0.740323763
East Asia-Central and Eastern Europe                             0.056799948 -0.585718740  0.699318636
Latin America and Caribbean-Central and Eastern Europe           0.323735705 -0.096891512  0.744362923
Middle East and North Africa-Central and Eastern Europe         -0.430891146 -0.877130871  0.015348579
North America and ANZ-Central and Eastern Europe                 1.515474975  0.755236611  2.275713339
South Asia-Central and Eastern Europe                           -1.182607216 -1.787296941 -0.577917492
Southeast Asia-Central and Eastern Europe                       -0.274683310 -0.824903442  0.275536822
Sub-Saharan Africa-Central and Eastern Europe                   -1.274555134 -1.639761831 -0.909348437
Western Europe-Central and Eastern Europe                        1.241169036  0.820541818  1.661796253
East Asia-Commonwealth of Independent States                     0.008489990 -0.843907373  0.860887353
Latin America and Caribbean-Commonwealth of Independent States   0.275425747 -0.425058554  0.975910049
Middle East and North Africa-Commonwealth of Independent States -0.479201104 -1.195358135  0.236955927
North America and ANZ-Commonwealth of Independent States         1.467165017  0.522859256  2.411470778
South Asia-Commonwealth of Independent States                   -1.230917175 -2.055174819 -0.406659531
Southeast Asia-Commonwealth of Independent States               -0.322993268 -1.108164074  0.462177538
Sub-Saharan Africa-Commonwealth of Independent States           -1.322865093 -1.991540599 -0.654189587
Western Europe-Commonwealth of Independent States                1.192859077  0.492374776  1.893343379
Latin America and Caribbean-East Asia                            0.266935757 -0.384697126  0.918568641
Middle East and North Africa-East Asia                          -0.487691094 -1.156143075  0.180760887
North America and ANZ-East Asia                                  1.458675027  0.550016384  2.367333669
South Asia-East Asia                                            -1.239407165 -2.022572423 -0.456241907
Southeast Asia-East Asia                                        -0.331483258 -1.073399933  0.410433417
Sub-Saharan Africa-East Asia                                    -1.331355083 -1.948667070 -0.714043095
Western Europe-East Asia                                         1.184369087  0.532736204  1.836001971
Middle East and North Africa-Latin America and Caribbean        -0.754626851 -1.213892616 -0.295361086
North America and ANZ-Latin America and Caribbean                1.191739270  0.423782552  1.959695987
South Asia-Latin America and Caribbean                          -1.506342922 -2.120708296 -0.891977548
Southeast Asia-Latin America and Caribbean                      -0.598419015 -1.159255305 -0.037582725
Sub-Saharan Africa-Latin America and Caribbean                  -1.598290840 -1.979304071 -1.217277609
Western Europe-Latin America and Caribbean                       0.917433330  0.483011408  1.351855252
North America and ANZ-Middle East and North Africa               1.946366121  1.164087298  2.728644943
South Asia-Middle East and North Africa                         -0.751716071 -1.383892826 -0.119539316
Southeast Asia-Middle East and North Africa                      0.156207836 -0.424085177  0.736500848
Sub-Saharan Africa-Middle East and North Africa                 -0.843663989 -1.252777320 -0.434550657
Western Europe-Middle East and North Africa                      1.672060181  1.212794416  2.131325946
South Asia-North America and ANZ                                -2.698082192 -3.580397287 -1.815767097
Southeast Asia-North America and ANZ                            -1.790158285 -2.636073445 -0.944243125
Sub-Saharan Africa-North America and ANZ                        -2.790030110 -3.529087750 -2.050972469
Western Europe-North America and ANZ                            -0.274305940 -1.042262657  0.493650778
Southeast Asia-South Asia                                        0.907923907  0.198515878  1.617331935
Sub-Saharan Africa-South Asia                                   -0.091947918 -0.669783092  0.485887256
Western Europe-South Asia                                        2.423776252  1.809410878  3.038141626
Sub-Saharan Africa-Southeast Asia                               -0.999871825 -1.520435000 -0.479308649
Western Europe-Southeast Asia                                    1.515852345  0.955016055  2.076688635
Western Europe-Sub-Saharan Africa                                2.515724170  2.134710939  2.896737401
pval
Commonwealth of Independent States-Central and Eastern Europe                0.89044
East Asia-Central and Eastern Europe                                         0.86153
Latin America and Caribbean-Central and Eastern Europe                       0.13038
Middle East and North Africa-Central and Eastern Europe                      0.05830 .
North America and ANZ-Central and Eastern Europe                             0.00013 ***
South Asia-Central and Eastern Europe                                        0.00017 ***
Southeast Asia-Central and Eastern Europe                                    0.32540
Sub-Saharan Africa-Central and Eastern Europe                      0.000000000156875 ***
Western Europe-Central and Eastern Europe                          0.000000034884740 ***
East Asia-Commonwealth of Independent States                                 0.98432
Latin America and Caribbean-Commonwealth of Independent States               0.43831
Middle East and North Africa-Commonwealth of Independent States              0.18806
North America and ANZ-Commonwealth of Independent States                     0.00255 **
```

```
South Asia-Commonwealth of Independent States                                0.00369 **
Southeast Asia-Commonwealth of Independent States                            0.41749
Sub-Saharan Africa-Commonwealth of Independent States                        0.00014 ***
Western Europe-Commonwealth of Independent States                            0.00098 ***
Latin America and Caribbean-East Asia                                        0.41944
Middle East and North Africa-East Asia                                       0.15144
North America and ANZ-East Asia                                              0.00185 **
South Asia-East Asia                                                         0.00213 **
Southeast Asia-East Asia                                                     0.37863
Sub-Saharan Africa-East Asia                                       0.000036415915935 ***
Western Europe-East Asia                                                     0.00045 ***
Middle East and North Africa-Latin America and Caribbean                     0.00145 **
North America and ANZ-Latin America and Caribbean                            0.00258 **
South Asia-Latin America and Caribbean                             0.000003232231723 ***
Southeast Asia-Latin America and Caribbean                                   0.03667 *
Sub-Saharan Africa-Latin America and Caribbean                     0.000000000000074 ***
Western Europe-Latin America and Caribbean                         0.000051647046219 ***
North America and ANZ-Middle East and North Africa                 0.000002368250056 ***
South Asia-Middle East and North Africa                                      0.02012 *
Southeast Asia-Middle East and North Africa                                  0.59548
Sub-Saharan Africa-Middle East and North Africa                    0.000075601986413 ***
Western Europe-Middle East and North Africa                        0.000000000032034 ***
South Asia-North America and ANZ                                   0.000000012376285 ***
Southeast Asia-North America and ANZ                               0.000049921601323 ***
Sub-Saharan Africa-North America and ANZ                           0.000000000007587 ***
Western Europe-North America and ANZ                                         0.48130
Southeast Asia-South Asia                                                    0.01250 *
Sub-Saharan Africa-South Asia                                                0.75357
Western Europe-South Asia                                          0.000000000001190 ***
Sub-Saharan Africa-Southeast Asia                                            0.00022 ***
Western Europe-Southeast Asia                                      0.000000352904692 ***
Western Europe-Sub-Saharan Africa                               < 0.0000000000000002 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2.1 Q2 Box plots

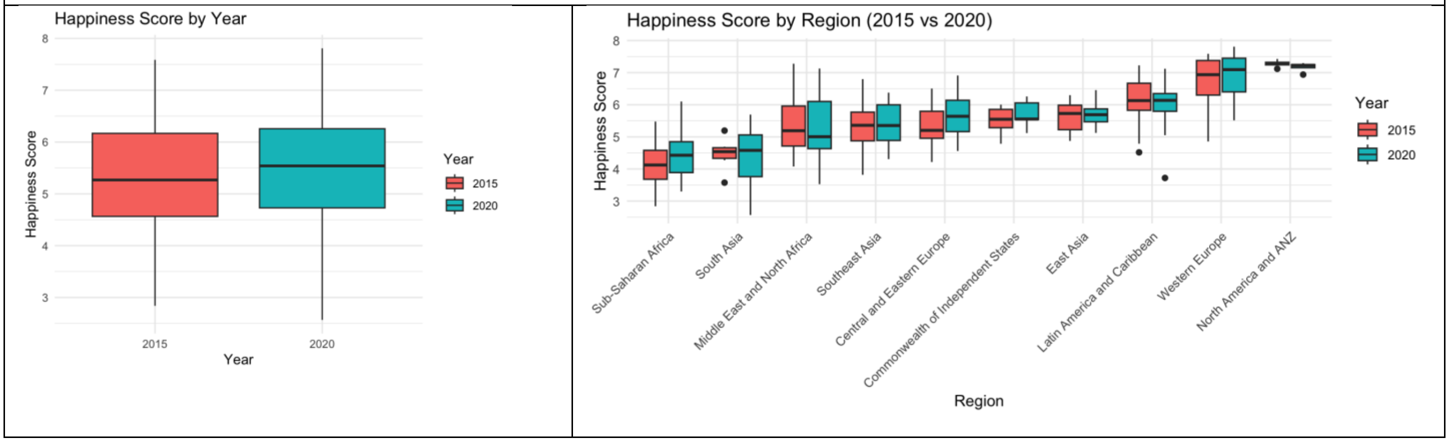


Figure 2.2 QQ Plots

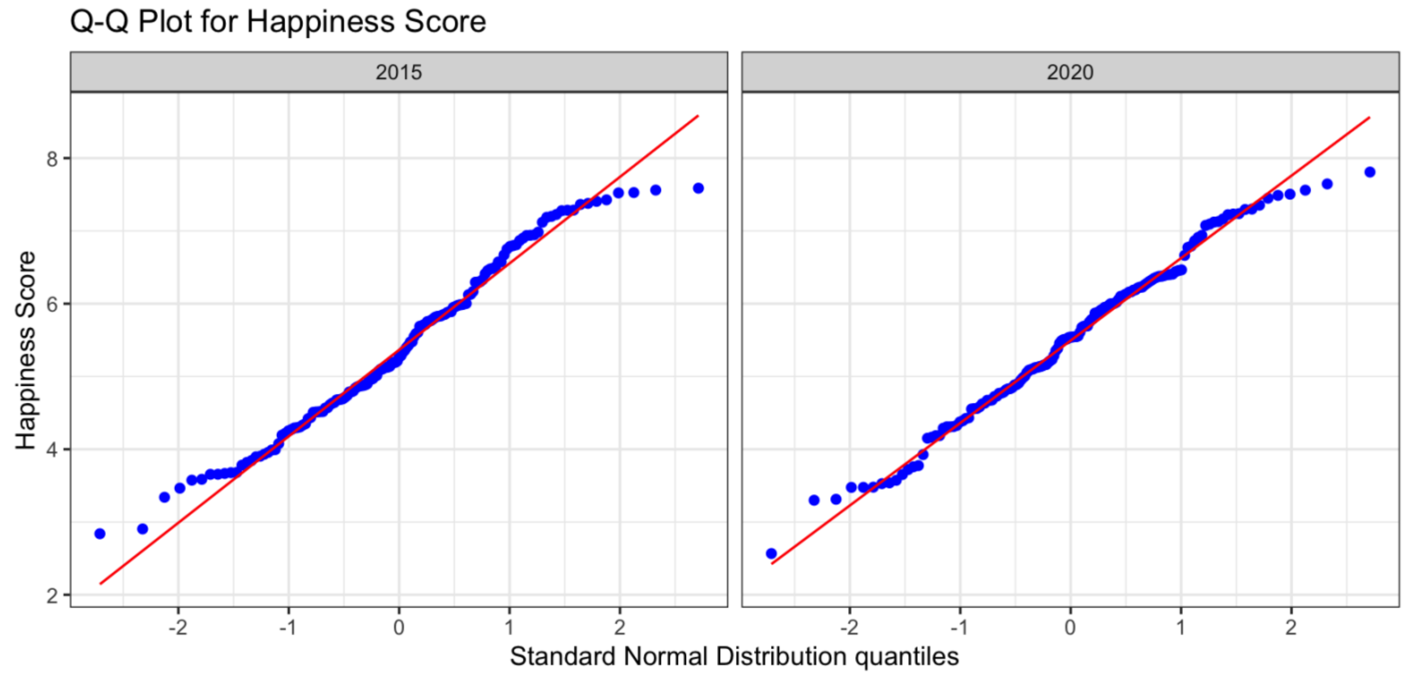


Figure 2.3 Q2 Histograms

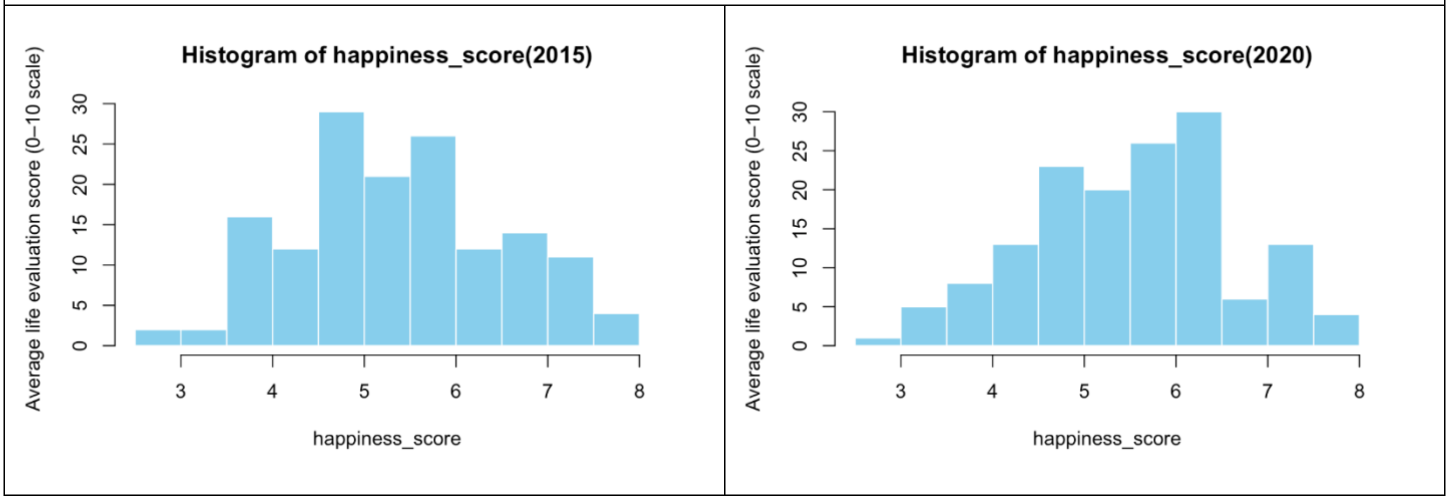


Figure 2.4 Q2 X's correlations

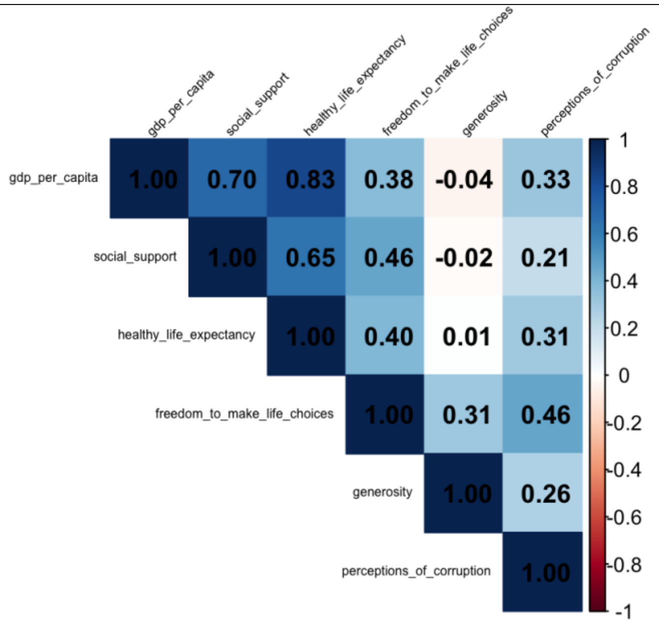
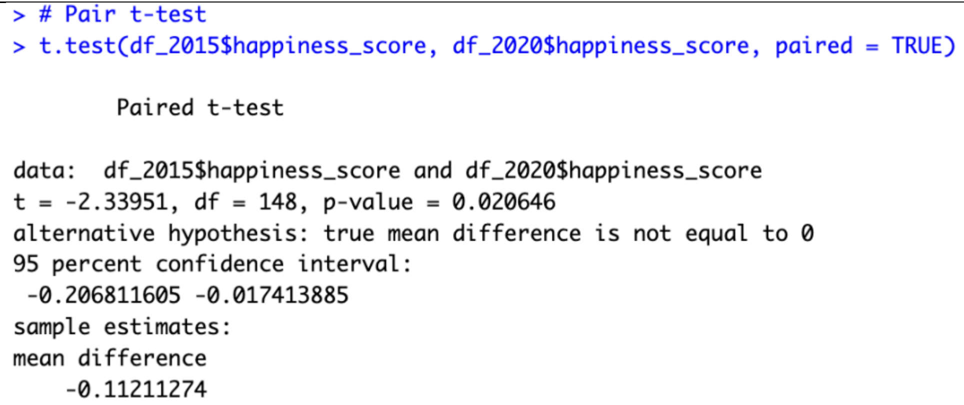


Figure 2.5 Q2 Pair t-test



```
> # Pair t-test
> t.test(df_2015$happiness_score, df_2020$happiness_score, paired = TRUE)

	Paired t-test

data:  df_2015$happiness_score and df_2020$happiness_score
t = -2.33951, df = 148, p-value = 0.020646
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.206811605 -0.017413885
sample estimates:
mean difference
    -0.11211274
```

Figure 3.1 Assumptions

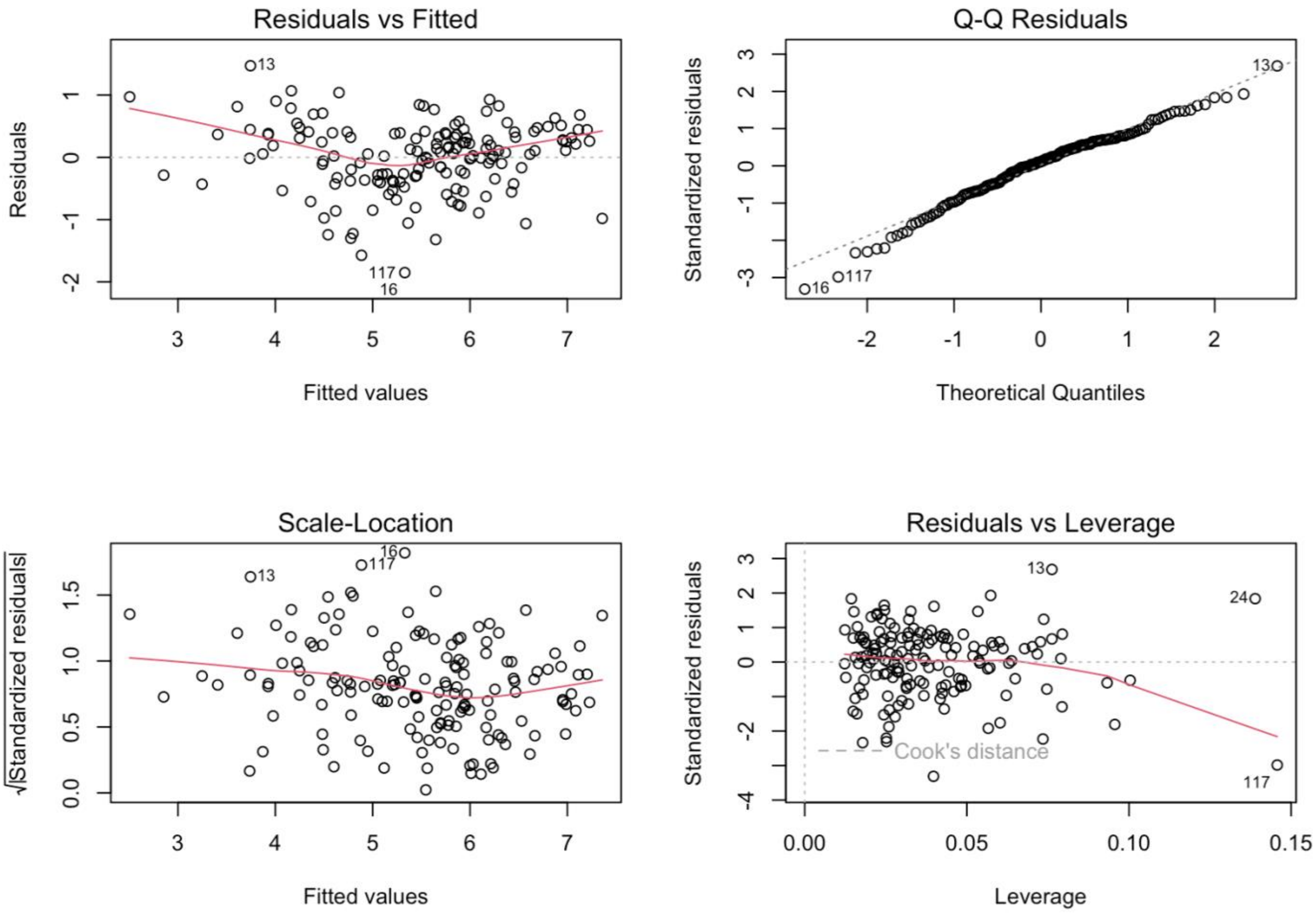


Figure 3.2 Multicollinearity

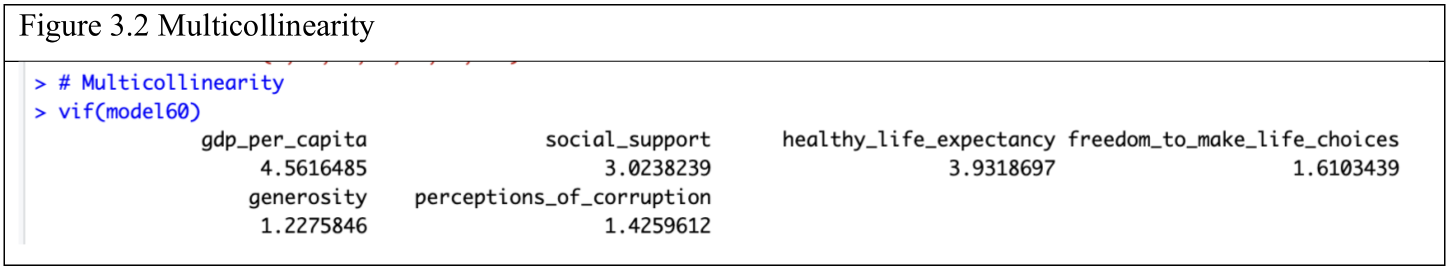
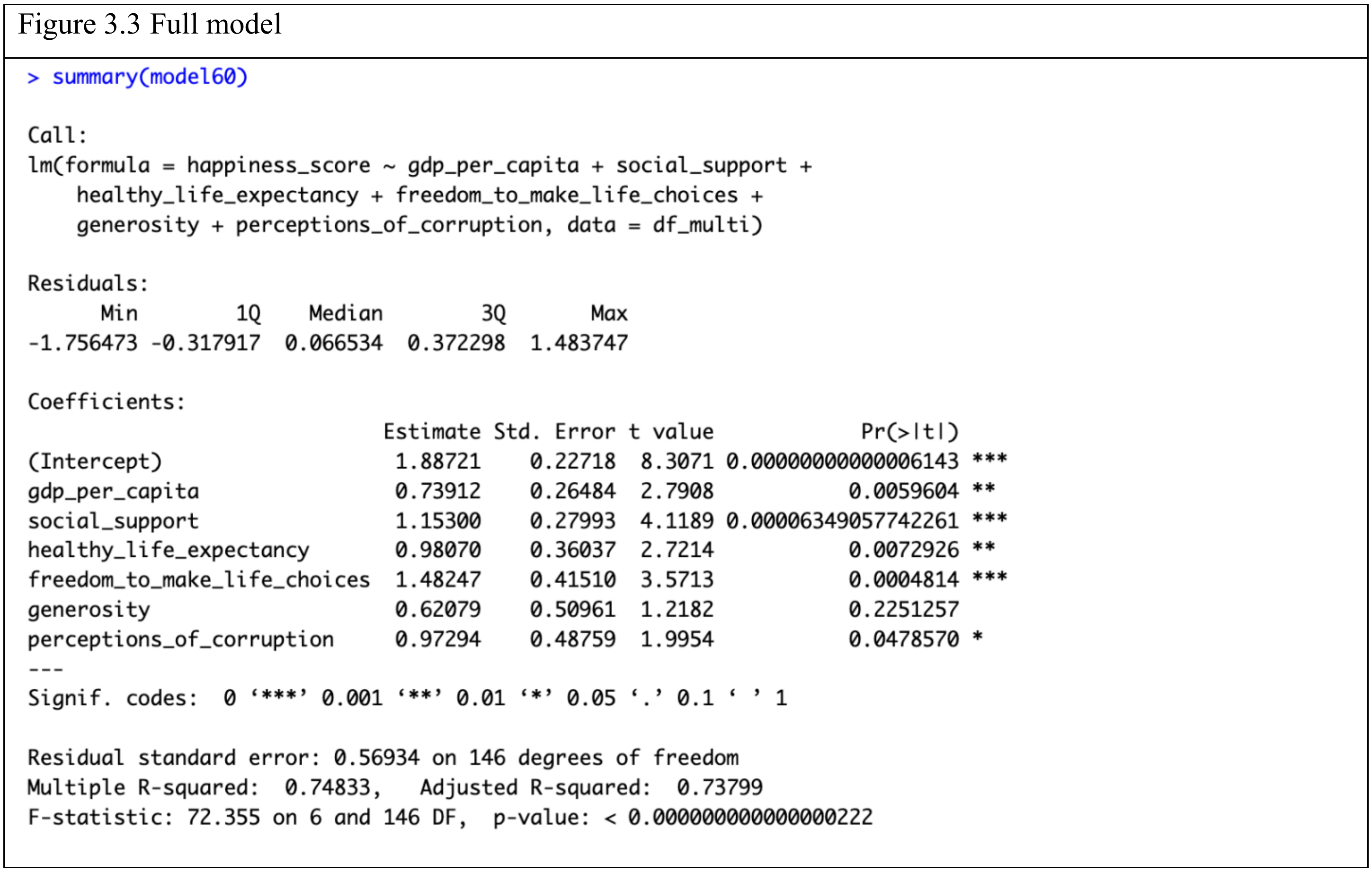
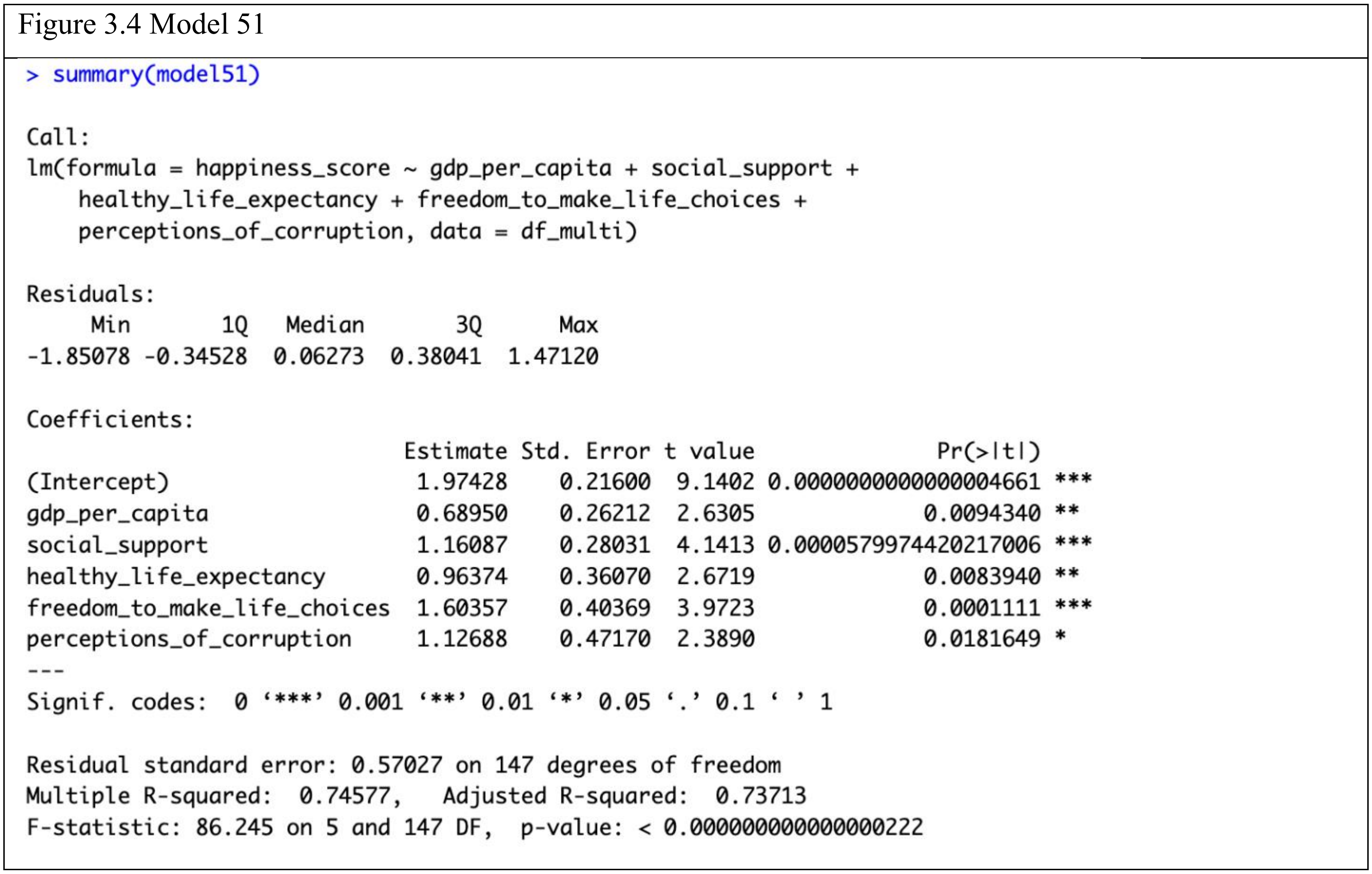

```
> # Multicollinearity
> vif(model60)
           gdp_per_capita            social_support     healthy_life_expectancy freedom_to_make_life_choices
                4.5616485                 3.0238239                   3.9318697                    1.6103439
               generosity perceptions_of_corruption
                1.2275846                 1.4259612
```

Figure 3.3 Full model



```
> summary(model60)

Call:
lm(formula = happiness_score ~ gdp_per_capita + social_support +
    healthy_life_expectancy + freedom_to_make_life_choices +
    generosity + perceptions_of_corruption, data = df_multi)

Residuals:
      Min        1Q    Median        3Q       Max
-1.756473 -0.317917  0.066534  0.372298  1.483747

Coefficients:
                             Estimate Std. Error t value           Pr(>|t|)
(Intercept)                   1.88721    0.22718  8.3071 0.00000000000006143 ***
gdp_per_capita                0.73912    0.26484  2.7908           0.0059604 **
social_support                1.15300    0.27993  4.1189 0.00006349057742261 ***
healthy_life_expectancy       0.98070    0.36037  2.7214           0.0072926 **
freedom_to_make_life_choices  1.48247    0.41510  3.5713           0.0004814 ***
generosity                    0.62079    0.50961  1.2182           0.2251257
perceptions_of_corruption     0.97294    0.48759  1.9954           0.0478570 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.56934 on 146 degrees of freedom
Multiple R-squared:  0.74833,   Adjusted R-squared:  0.73799
F-statistic: 72.355 on 6 and 146 DF,  p-value: < 0.0000000000000002220
```

Figure 3.4 Model 51



```
> summary(model51)

Call:
lm(formula = happiness_score ~ gdp_per_capita + social_support +
    healthy_life_expectancy + freedom_to_make_life_choices +
    perceptions_of_corruption, data = df_multi)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85078 -0.34528  0.06273  0.38041  1.47120

Coefficients:
                             Estimate Std. Error t value             Pr(>|t|)
(Intercept)                   1.97428    0.21600  9.1402 0.0000000000000004661 ***
gdp_per_capita                0.68950    0.26212  2.6305             0.0094340 **
social_support                1.16087    0.28031  4.1413 0.0000579974420217006 ***
healthy_life_expectancy       0.96374    0.36070  2.6719             0.0083940 **
freedom_to_make_life_choices  1.60357    0.40369  3.9723             0.0001111 ***
perceptions_of_corruption     1.12688    0.47170  2.3890             0.0181649 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.57027 on 147 degrees of freedom
Multiple R-squared:  0.74577,   Adjusted R-squared:  0.73713
F-statistic: 86.245 on 5 and 147 DF,  p-value: < 0.0000000000000002220
```