Project-2

Saniya Abushakimova

5/1/2021

Introduction

Aim

The purpose of this project is to analyze what factors influence the level of life satisfaction of people living in developing countries. The project particularly tries to give an answer to the following research question: *"What did affect the Life Satisfaction rate of people from developing countries in 2016?"*.

Dataset description

The dataset was collected manually by combining multiple charts from Our World in Data (https://ourworldindata.org/charts). The original charts contained a lot of information, however for the purpose of this project, only developing countries in 2016 were taken into consideration. More details on the data preprocessing part can be found in */data/data_preprocessing.ipynb*.

The dataset includes 84 observations, which corresponds to 84 developing countries, and 13 different numerical variables.

Response variable (Y):

`life_satisfaction` - life satisfaction rate. It illustrates the average of survey responses to the 'Cantril Ladder' question, in which the best possible life is rated as 10, and the worst one as 0. [Source (https://ourworldindata.org/grapher/gdp-vs-happiness)]

Predictor variables (X):

`GDP` - Gross Domestic Product. It is measured in $ and indicates how good is the economy of a country. [Source (https://ourworldindata.org/grapher/gdp-vs-happiness)]

`GHI` - Global Hunger Index. It measures the rate of hunger by country based on 4 key indicators: undernourishment, child wasting, child stunting, and child mortality.The scale is from 0 (no hunger) to 100 (severe hunger). [Source (https://ourworldindata.org/grapher/global-hunger-index?tab=table)]

`life_expectancy` - life expectancy in years. [Source (https://ourworldindata.org/grapher/life-satisfaction-vs-life-expectancy)]

`unemployment` - unemployment rate. It is a share of the labor force that is unemployed and is given in %. [Source (https://ourworldindata.org/grapher/unemployment-rate?tab=table)]

`HDI` - Human Development Index. It measures the average achievement of human development in three areas: a long and healthy life, being knowledgeable, and a decent standard of living. The scale is from 0 (very low) to 1 (very high). [Source (https://ourworldindata.org/human-development-index)]

`corruption` - corruption rate. It measures the level of corruption by country. The scale is from 100 (no corruption) to 0 (very high corruption rate). [Source (https://ourworldindata.org/grapher/ti-corruption-perception-index?tab=table&time=2012..2017)]

`depression` - percentage of people with depressive disorders. It is given in %. [Source (https://ourworldindata.org/grapher/depression-vs-self-reported-life-satisfaction?tab=table)]

`obesity` - share of people (aged 18 and above) who are obese (BMI>30). It is given in %. [Source (https://ourworldindata.org/grapher/share-of-adults-defined-as-obese?tab=table)]

`suicide` - suicide mortality rate. It measures the number of siucide deaths in a year per 100,000 population.

[Source (https://ourworldindata.org/grapher/suicide-rate?tab=table)]
`fruit_consumption` - average supply of fruit across the population. It is measured in kilograms per person per year. [Source (https://ourworldindata.org/grapher/fruit-consumption-per-capita)]
`veg_consumption` - average supply of vegetables across the population. It is measured in kilograms per person per year. [Source (https://ourworldindata.org/grapher/vegetable-consumption-per-capita)]
`birth_rate` - the number of live births occurring during the year, per 1,000 people. [Source (https://ourworldindata.org/grapher/crude-birth-rate?country=~OWID_WRL)]
`under_five_deaths` - under-five children mortality rate. It indicates deaths under age age 5 per 1,000 live births, and is given in %. [Source (https://ourworldindata.org/grapher/correlation-between-child-mortality-and-mean-years-of-schooling-for-those-aged-15-and-older?tab=table&time=2016)]

This is an **observational study**, which means no preliminary research on topic was done. All predictor variables were chosen based on the assumption that they might somehow be related to the life satisfaction rate.

```
# Loading all the necessary packages
library(rmarkdown)
library(car)
library(corrplot)
library(latex2exp)
library(olsrr)
library(lmtest)
library(leaps)
```

The dataset is shown below:

```
# Loading and printing the data
data <- data.frame(read.csv(file ='~/Desktop/Project-2/data/final_dataset.csv'))
paged_table(data)
```

Ind…	Country	Y…	GDP	life_satisfaction	GHI	life_expectancy
<int>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
0	Afghanistan	2016	1802.6956	4.220169	34.8	63.763
1	Albania	2016	11356.3413	4.511101	11.9	78.194
2	Algeria	2016	13921.1800	5.340854	8.7	76.298
3	Argentina	2016	18584.5800	6.427221	5.0	76.221
4	Armenia	2016	8190.2393	4.325472	8.7	74.640
5	Azerbaijan	2016	16001.3234	5.303895	9.8	72.493
6	Bangladesh	2016	3319.3528	4.556141	27.1	71.785
7	Belarus	2016	16773.1936	5.177899	5.0	74.031
8	Benin	2016	2009.6183	4.007358	23.2	60.885
9	Bosnia and Herzegovina	2016	11337.5703	5.180865	5.0	76.998

1-10 of 84 rows | 1-7 of 17 columns Previous **1** 2 3 4 5 6 … 9 Next

```
# Removing unnecessary columns
data_reduced_1 <- subset(data, select = -c(Index,Country,Year))
```

Exploratory Data Analysis

Let's start the exploratory data analysis from observing the **statistical summary**. It will help us to understand the data in general terms.

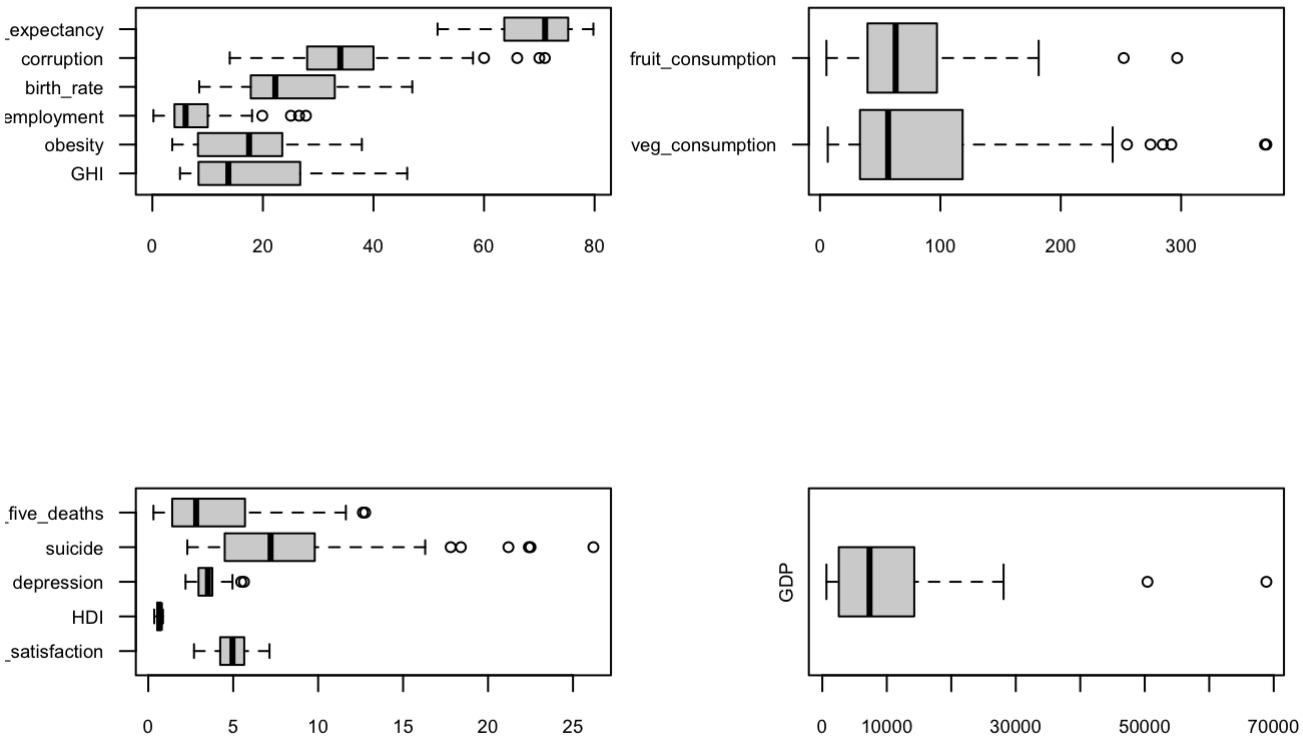
```
summary(data_reduced_1)
```

```
##       GDP          life_satisfaction      GHI         life_expectancy
##  Min.   :  642.9   Min.   :2.693     Min.   : 5.000   Min.   :51.59
##  1st Qu.: 2647.1   1st Qu.:4.234     1st Qu.: 8.425   1st Qu.:63.71
##  Median : 7327.3   Median :4.957     Median :13.750   Median :71.04
##  Mean   : 9950.4   Mean   :4.949     Mean   :17.755   Mean   :69.32
##  3rd Qu.:14190.4   3rd Qu.:5.649     3rd Qu.:26.675   3rd Qu.:75.23
##  Max.   :68861.8   Max.   :7.136     Max.   :46.100   Max.   :79.78
##   unemployment         HDI           corruption      depression
##  Min.   : 0.198   Min.   :0.3510   Min.   :14.00   Min.   :2.194
##  1st Qu.: 4.060   1st Qu.:0.5275   1st Qu.:28.00   1st Qu.:2.958
##  Median : 6.009   Median :0.6850   Median :34.00   Median :3.512
##  Mean   : 7.640   Mean   :0.6477   Mean   :34.98   Mean   :3.410
##  3rd Qu.: 9.932   3rd Qu.:0.7530   3rd Qu.:40.00   3rd Qu.:3.760
##  Max.   :27.782   Max.   :0.8680   Max.   :71.00   Max.   :5.636
##     obesity          suicide       fruit_consumption veg_consumption
##  Min.   : 3.600   Min.   : 2.300   Min.   :  5.36    Min.   :  6.46
##  1st Qu.: 8.325   1st Qu.: 4.550   1st Qu.: 39.48    1st Qu.: 33.31
##  Median :17.500   Median : 7.200   Median : 62.83    Median : 56.59
##  Mean   :16.848   Mean   : 8.132   Mean   : 72.77    Mean   : 88.26
##  3rd Qu.:23.400   3rd Qu.: 9.750   3rd Qu.: 96.87    3rd Qu.:118.26
##  Max.   :37.900   Max.   :26.200   Max.   :296.71    Max.   :370.82
##    birth_rate     under_five_deaths
##  Min.   : 8.477   Min.   : 0.3054
##  1st Qu.:17.832   1st Qu.: 1.4198
##  Median :22.239   Median : 2.8192
##  Mean   :24.392   Mean   : 4.0228
##  3rd Qu.:32.932   3rd Qu.: 5.6465
##  Max.   :47.017   Max.   :12.7672
```

From the above table we can observe that:

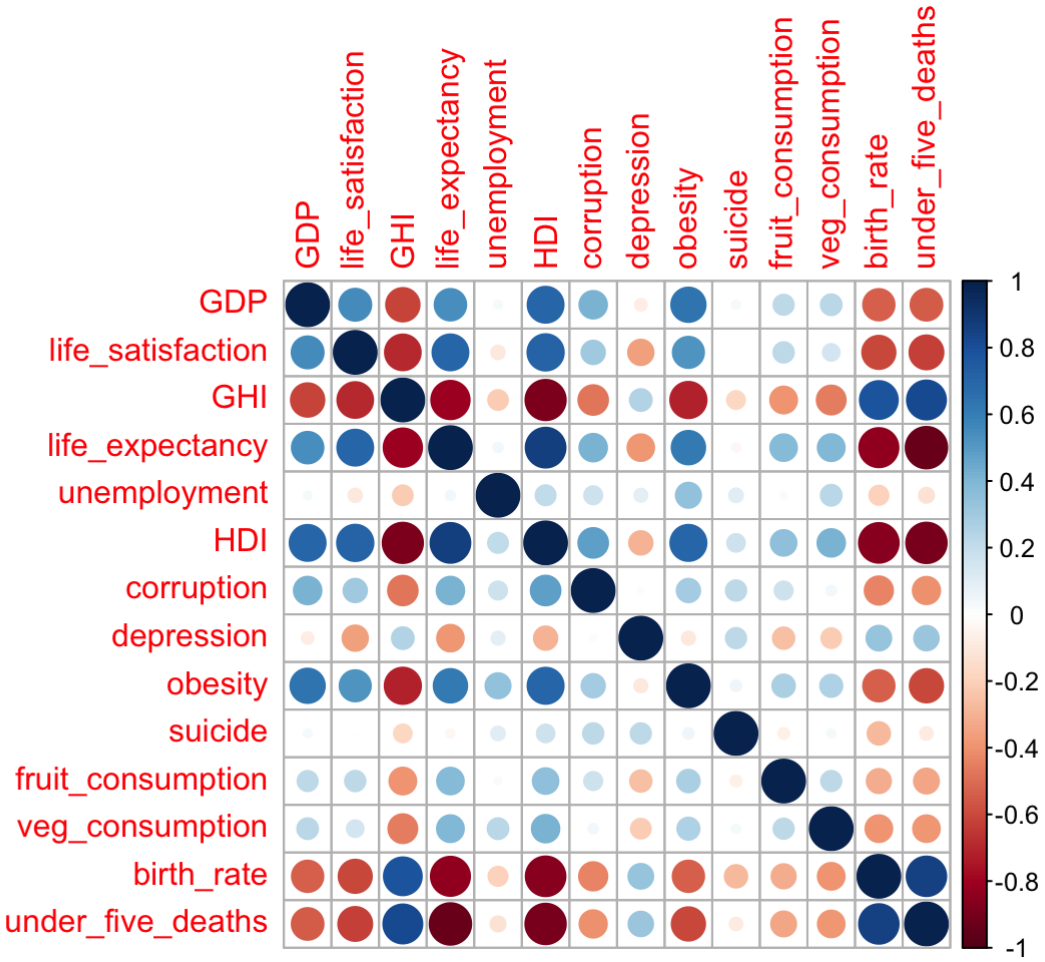
1. Low variability - `life_satisfaction`, `HDI`, `depression`.
2. Moderate variability - `GHI`, `life_expectancy`, `unemployment`, `corruption`, `obesity`, `suicide`, `under_five_deaths`, `birth_rate`.
3. High variability - `GDP`, `veg_consumption`, `fruit consumption`.
4. Potential outliers can be in `GDP`, `GHI`, `obesity`, `suicide`, `fruit_consumption`, `veg_consumption`, `birth_rate`, `under_five_deaths`, `unemployment` because the mean and the median of these features are not close to each other. Let's verify 4 by looking at the **box plots**.

```
# Plotting the box plots
par(mfrow=c(2,2))
par(cex.axis=0.7)
boxplot(data_reduced_1[c("GHI", "obesity", "unemployment", "birth_rate", "corruption","l
ife_expectancy")],las=1,horizontal = TRUE)
boxplot(data_reduced_1[c("veg_consumption","fruit_consumption")],las=1,horizontal = TRUE
)
boxplot(data_reduced_1[c("life_satisfaction","HDI","depression","suicide","under_five_de
aths")],las=1,horizontal = TRUE)
boxplot(data_reduced_1$GDP,horizontal = TRUE)
title(ylab="GDP", line=0.3, cex.lab=0.7)
```



As it can be seen from the above box plots, `corruption` , `unemployment` , `fruit_consumption` , `veg_consumption` , `under_five_deaths` , `suicide` , `depression` , `GDP` have outliers. We might consider to remove some of them in future if they influence the model significantly. Now, it is worth to investigate the **correlation** between variables, becuase it might give us a clue whether multicollinearity may be present in the model.

```
# Plotting the correlation matrix
corr.matrix <- cor(data_reduced_1)
corrplot(corr.matrix, method="circle")
```



The correlation matrix shows that there are many variables that are highly correlated with each other. This may serve as a warning for us to carefully investigate multicollinearity in the model.

Feature/Model Selection

Let's start our model selection with simple additive model that contains all variables, and let's observe the p-values for the individual F-tests.

```
# Fitting an additive model containing all 13 variables
m_1 <- lm(life_satisfaction~., data = data_reduced_1)
summary(m_1)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ ., data = data_reduced_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00772 -0.34054 -0.00039  0.31092  1.09980
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.891e+00  2.783e+00  -1.398 0.166485
## GDP                6.952e-06  1.007e-05   0.690 0.492383
## GHI               -4.468e-02  1.341e-02  -3.331 0.001385 **
## life_expectancy    1.039e-01  3.282e-02   3.167 0.002281 **
## unemployment      -2.288e-02  1.405e-02  -1.629 0.107905
## HDI                4.314e+00  1.563e+00   2.761 0.007356 **
## corruption        -1.142e-02  6.785e-03  -1.684 0.096681 .
## depression        -1.535e-01  1.118e-01  -1.373 0.174287
## obesity           -7.515e-03  1.203e-02  -0.624 0.534340
## suicide            1.391e-03  1.667e-02   0.083 0.933726
## fruit_consumption -2.392e-03  1.289e-03  -1.855 0.067850 .
## veg_consumption   -2.784e-03  8.936e-04  -3.116 0.002659 **
## birth_rate         9.414e-03  1.602e-02   0.588 0.558600
## under_five_deaths  2.409e-01  6.571e-02   3.667 0.000475 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.541 on 70 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.6714
## F-statistic: 14.04 on 13 and 70 DF,  p-value: 9.737e-15
```

According to thee individual F-tests, some of the variables seem to be significant and some of them not, however it seems that these significance tests may be misleading because of the multicollinearity that may be present in our model. A systematic way to detect multicollinearity would be to look at the **variance inflation factor (VIF)** scores.

```
vif(m_1)
```

```
##               GDP               GHI   life_expectancy      unemployment
##          3.122587          6.176820         15.582687          1.871906
##               HDI        corruption        depression           obesity
##         12.215459          1.727845          1.473524          3.520205
##           suicide fruit_consumption   veg_consumption        birth_rate
##          1.940674          1.257204          1.499767          6.479826
## under_five_deaths
##         12.816918
```

As we can see there are some severe multicollinearities present in the model. It is important to handle them because otherwise it will be difficult to identify which variables are truly significant to the model. We will handle multicollinearity by dropping the variables with the most severe VIF scores one at a time. The most severe ones

are considered to be those larger than 10. We will also try to balance VIF's in such a way that they all be under some threshold, which is usually 5.

Handling Multicollinearity

Let's remove `life_expectancy` first becuase it has the largest VIF value and see how the removal will affect other variables.

```
# Removing 'life_expectancy'
data_reduced_2 <- subset(data_reduced_1, select = -c(life_expectancy))
m_2 <- lm(life_satisfaction~., data = data_reduced_2)
vif(m_2)
```

```
##               GDP               GHI      unemployment               HDI
##          2.894493          6.131150          1.497218         12.003710
##        corruption        depression           obesity           suicide
##          1.590836          1.472120          3.116669          1.460010
## fruit_consumption   veg_consumption        birth_rate under_five_deaths
##          1.246117          1.450683          5.516026          7.205050
```

As it can be seen from the values there is still multicollinearity present, therefore let's proceed with removing variables with the highest VIF score.

```
# Removing 'HDI'
data_reduced_3 <- subset(data_reduced_2, select = -c(HDI))
m_3 <- lm(life_satisfaction~., data = data_reduced_3)
vif(m_3)
```

```
##               GDP               GHI      unemployment        corruption
##          2.306139          5.503059          1.449621          1.585471
##        depression           obesity           suicide fruit_consumption
##          1.435094          3.083689          1.434664          1.244591
##   veg_consumption        birth_rate under_five_deaths
##          1.450429          5.219266          5.544723
```

```
# Removing 'under_five_deaths'
data_reduced_4 <- subset(data_reduced_3, select = -c(under_five_deaths))
m_4 <- lm(life_satisfaction~., data = data_reduced_4)
vif(m_4)
```

```
##               GDP               GHI      unemployment        corruption
##          2.289892          4.735711          1.398716          1.584305
##        depression           obesity           suicide fruit_consumption
##          1.434344          3.035923          1.322525          1.242255
##   veg_consumption        birth_rate
##          1.450204          3.220751
```

After removing three variables with the highest multicollinearity, the resultant dataset that we will utilize further is `data_reduced_4` . It containst 10 variables and all of their VIF scores are below the threshold=5.
We will now proceed with finding the subset of variables that will explain `life_satisfaction` in the best way.

Best subset

```
# Finding the best subset of variables
subsets <-regsubsets(life_satisfaction~., nbest=3, data=data_reduced_4)
all_output <- summary(subsets)
with(all_output, round(cbind(which, adjr2, cp, bic),3))
```

```
##   (Intercept) GDP GHI unemployment corruption depression obesity suicide
## 1           1   0   1            0          0          0       0       0
## 1           1   0   0            0          0          0       0       0
## 1           1   1   0            0          0          0       0       0
## 2           1   0   1            1          0          0       0       0
## 2           1   0   1            0          0          0       0       0
## 2           1   0   1            0          0          1       0       0
## 3           1   0   1            1          0          0       0       0
## 3           1   0   1            1          0          1       0       0
## 3           1   1   1            1          0          0       0       0
## 4           1   0   1            1          0          1       0       0
## 4           1   1   1            1          0          1       0       0
## 4           1   1   1            0          0          1       0       0
## 5           1   1   1            1          0          1       0       0
## 5           1   0   1            1          0          1       1       0
## 5           1   0   1            1          0          1       0       0
## 6           1   1   1            1          0          1       0       0
## 6           1   1   1            1          1          1       0       0
## 6           1   1   1            1          0          1       0       0
## 7           1   1   1            1          0          1       0       1
## 7           1   1   1            1          0          1       1       0
## 7           1   1   1            1          1          1       0       0
## 8           1   1   1            1          0          1       0       1
## 8           1   1   1            1          0          1       1       0
## 8           1   1   1            1          1          1       0       0
##   fruit_consumption veg_consumption birth_rate adjr2     cp     bic
## 1                 0               0          0 0.479 24.675 -46.886
## 1                 0               0          1 0.354 49.702 -28.877
## 1                 0               0          0 0.302 60.139 -22.377
## 2                 0               0          0 0.537 13.796 -53.483
## 2                 0               1          0 0.508 19.683 -48.262
## 2                 0               0          0 0.505 20.275 -47.755
## 3                 0               1          0 0.554 11.396 -53.178
## 3                 0               0          0 0.552 11.841 -52.752
## 3                 0               0          0 0.551 12.028 -52.573
## 4                 0               1          0 0.576  8.083 -54.016
## 4                 0               0          0 0.572  8.898 -53.186
## 4                 0               1          0 0.570  9.195 -52.886
## 5                 0               1          0 0.595  5.382 -54.539
## 5                 0               1          0 0.583  7.565 -52.203
## 5                 1               1          0 0.578  8.637 -51.078
## 6                 1               1          0 0.597  6.001 -51.621
## 6                 0               1          0 0.593  6.706 -50.846
## 6                 0               1          1 0.593  6.787 -50.757
## 7                 1               1          0 0.596  7.268 -48.005
## 7                 1               1          0 0.595  7.369 -47.891
## 7                 1               1          0 0.595  7.386 -47.872
## 8                 1               1          1 0.597  8.024 -44.973
## 8                 1               1          1 0.594  8.610 -44.311
## 8                 1               1          1 0.594  8.623 -44.296
```

According to R_a^2, Mallows's C_p and BIC, the best subset is a combination of `veg_consumption` + `GDP` + `GHI` + `unemployment` + `depression` , row #5 (1). Let's fit this model.

```
# Fitting the model with the best subset of variables
best_m_1 <- lm(life_satisfaction~veg_consumption+GDP+GHI+unemployment+depression,data=da
ta_reduced_4)
summary(best_m_1)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ veg_consumption + GDP + GHI +
##     unemployment + depression, data = data_reduced_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18422 -0.38550 -0.00303  0.34593  1.34777
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.143e+00  4.203e-01  16.994  < 2e-16 ***
## veg_consumption -2.197e-03  9.316e-04  -2.358   0.0209 *
## GDP              1.769e-05  8.128e-06   2.177   0.0325 *
## GHI             -5.594e-02  8.650e-03  -6.466 8.01e-09 ***
## unemployment    -2.937e-02  1.213e-02  -2.421   0.0178 *
## depression      -2.811e-01  1.092e-01  -2.573   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6007 on 78 degrees of freedom
## Multiple R-squared:  0.6193, Adjusted R-squared:  0.5949
## F-statistic: 25.38 on 5 and 78 DF,  p-value: 4.205e-15
```

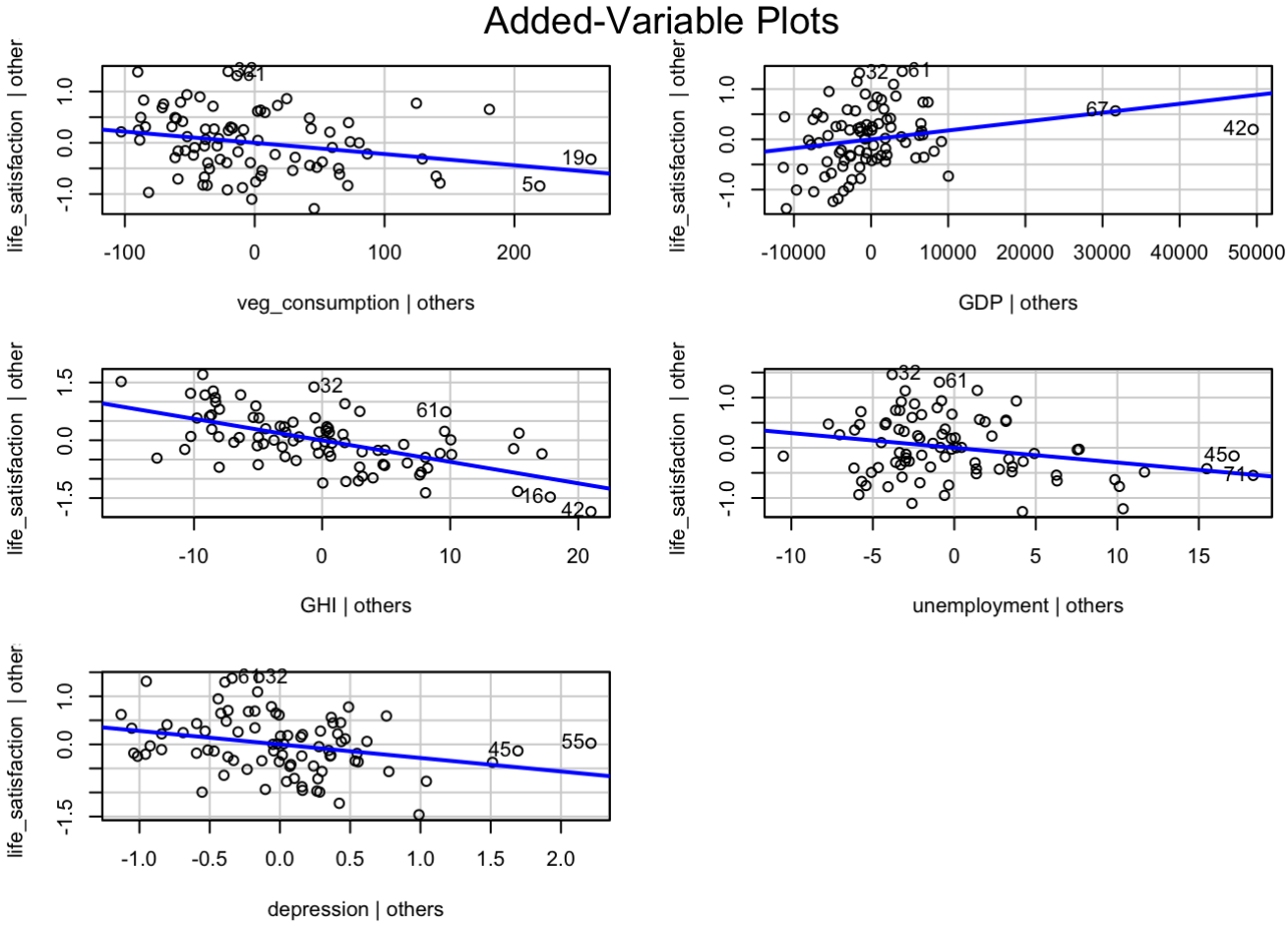
As it can be noticed all individual F-tests show that all variables present in this model are significant. We will now perform some model diagnostics to see what should be improved futher.

Model Diagnostics

1. Added-variable plots
2. Outliers/Influence points
3. Normality
4. Independence
5. Linearity/Homogeneity of variance

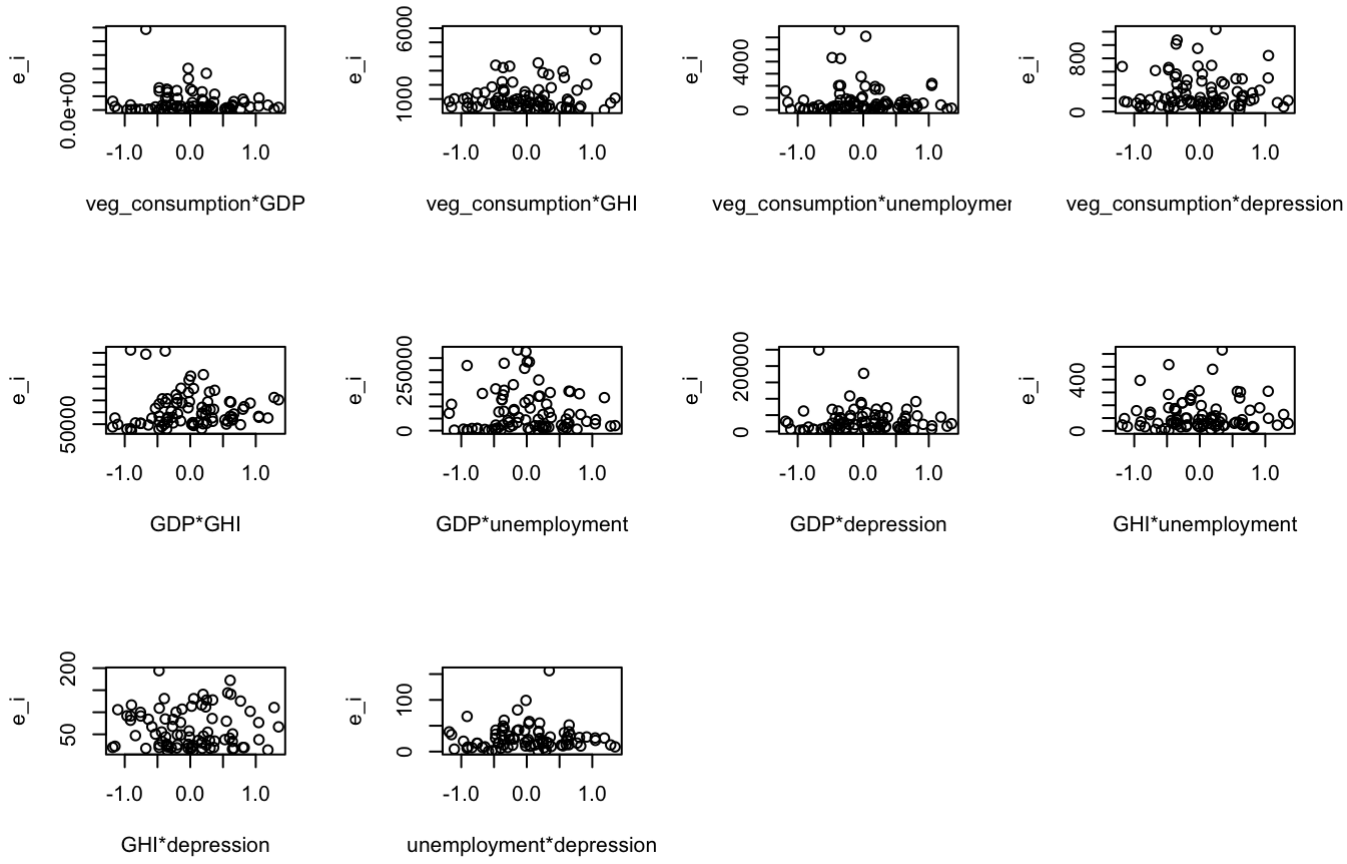
Added-variable plots

```
avPlots(best_m_1)
```



According to the Added-Variable plots, all variables seem to be significant to the model because the blue line has some slope. There is also no need for higher-order terms because the plots do not have any non-linear pattern. However we may need to have interaction terms. In order to investigate whether there is a need for interaction terms we can plot the residuals from our additive model vs each interaction term.

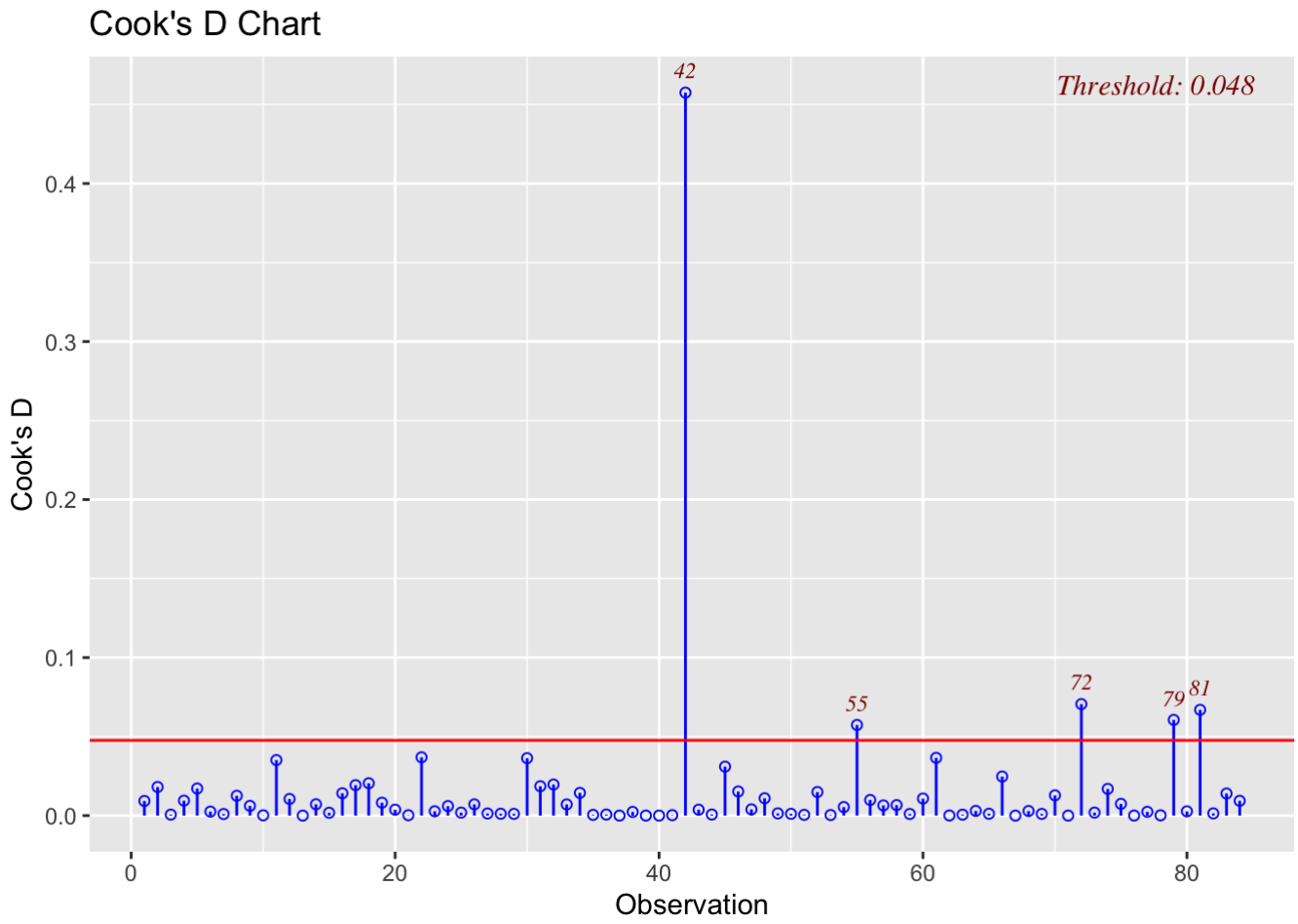
```
par(mfrow=c(3,4))
plot(best_m_1$residuals,data_reduced_4$veg_consumption*data_reduced_4$GDP,
     ylab="e_i",xlab="veg_consumption*GDP")
plot(best_m_1$residuals,data_reduced_4$veg_consumption*data_reduced_4$GHI,
     ylab="e_i",xlab="veg_consumption*GHI")
plot(best_m_1$residuals,data_reduced_4$veg_consumption*data_reduced_4$unemployment,
     ylab="e_i",xlab="veg_consumption*unemployment")
plot(best_m_1$residuals,data_reduced_4$veg_consumption*data_reduced_4$depression,
     ylab="e_i",xlab="veg_consumption*depression")
plot(best_m_1$residuals,data_reduced_4$GDP*data_reduced_4$GHI,
     ylab="e_i",xlab="GDP*GHI")
plot(best_m_1$residuals,data_reduced_4$GDP*data_reduced_4$unemployment,
     ylab="e_i",xlab="GDP*unemployment")
plot(best_m_1$residuals,data_reduced_4$GDP*data_reduced_4$depression,
     ylab="e_i",xlab="GDP*depression")
plot(best_m_1$residuals,data_reduced_4$GHI*data_reduced_4$unemployment,
     ylab="e_i",xlab="GHI*unemployment")
plot(best_m_1$residuals,data_reduced_4$GHI*data_reduced_4$depression,
     ylab="e_i",xlab="GHI*depression")
plot(best_m_1$residuals,data_reduced_4$unemployment*data_reduced_4$depression,
     ylab="e_i",xlab="unemployment*depression")
```



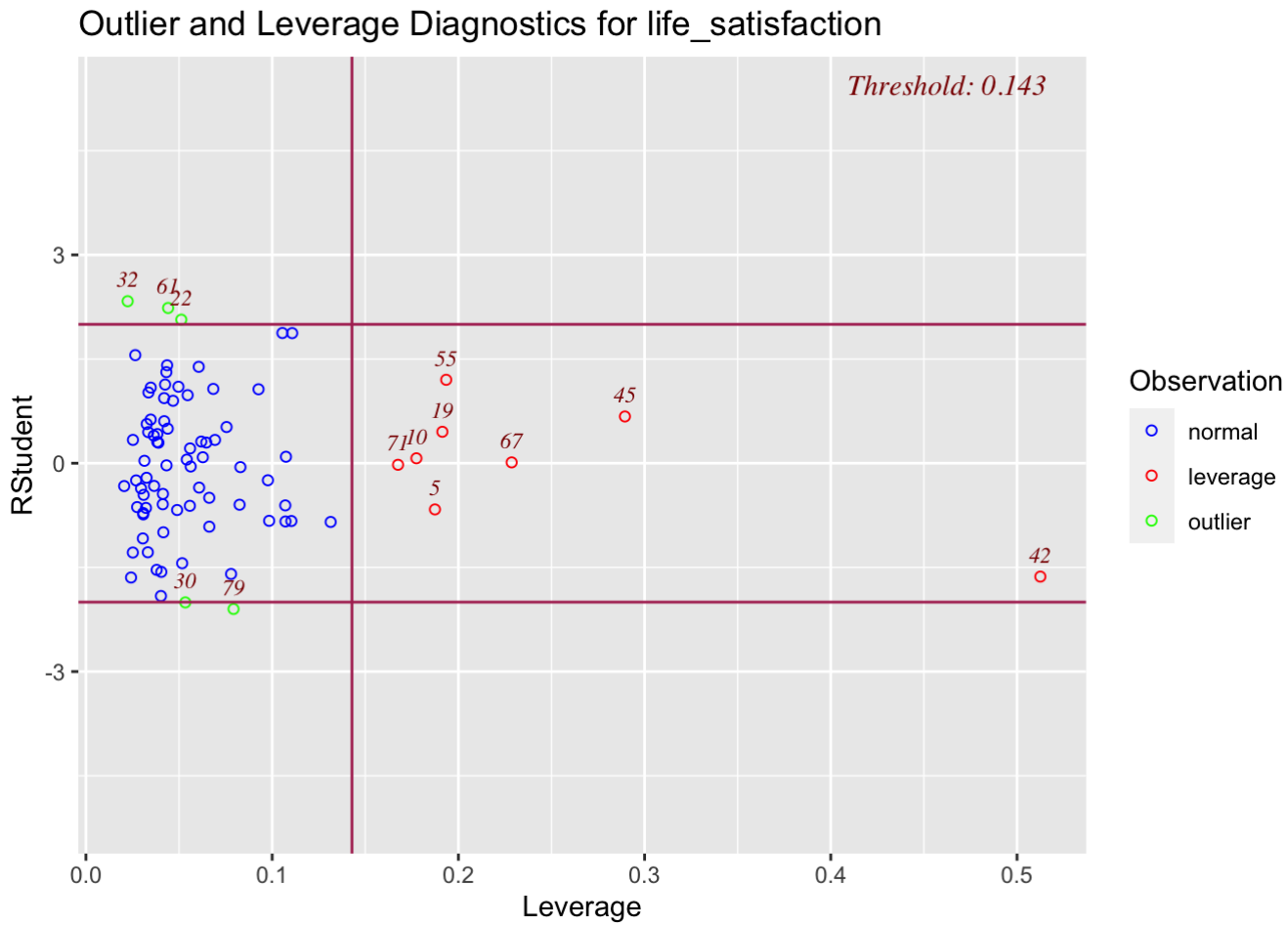
The above plots do not reveal any non-linear patter which suggests that our model does not need interaction terms either. We should now proceed with analyzing the outliers, and see if we can drop some of those that influence our model significantly.

Outliers

```
# Plotting Cook's distance chart
ols_plot_cooksd_chart(best_m_1)
```




```
# Plotting Studentized Residuals vs Leverage Plot
ols_plot_resid_lev(best_m_1)
```



From both of the plots, it seems that observation 42 is an influential point, therefore we can remove it from the dataset, and fit a new model.

```
# Removing observation 42 and fitting a new model
data_reduced_5 <- data_reduced_4[-c(42),]
best_m_2 <- lm(life_satisfaction~veg_consumption+GDP+GHI+unemployment+depression, data=d
ata_reduced_5)
```

Now let's compare the model with and without the influential point.

```
summary(best_m_1)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ veg_consumption + GDP + GHI +
##     unemployment + depression, data = data_reduced_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18422 -0.38550 -0.00303  0.34593  1.34777
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.143e+00  4.203e-01  16.994  < 2e-16 ***
## veg_consumption -2.197e-03  9.316e-04  -2.358   0.0209 *
## GDP              1.769e-05  8.128e-06   2.177   0.0325 *
## GHI             -5.594e-02  8.650e-03  -6.466 8.01e-09 ***
## unemployment    -2.937e-02  1.213e-02  -2.421   0.0178 *
## depression      -2.811e-01  1.092e-01  -2.573   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6007 on 78 degrees of freedom
## Multiple R-squared:  0.6193, Adjusted R-squared:  0.5949
## F-statistic: 25.38 on 5 and 78 DF,  p-value: 4.205e-15
```

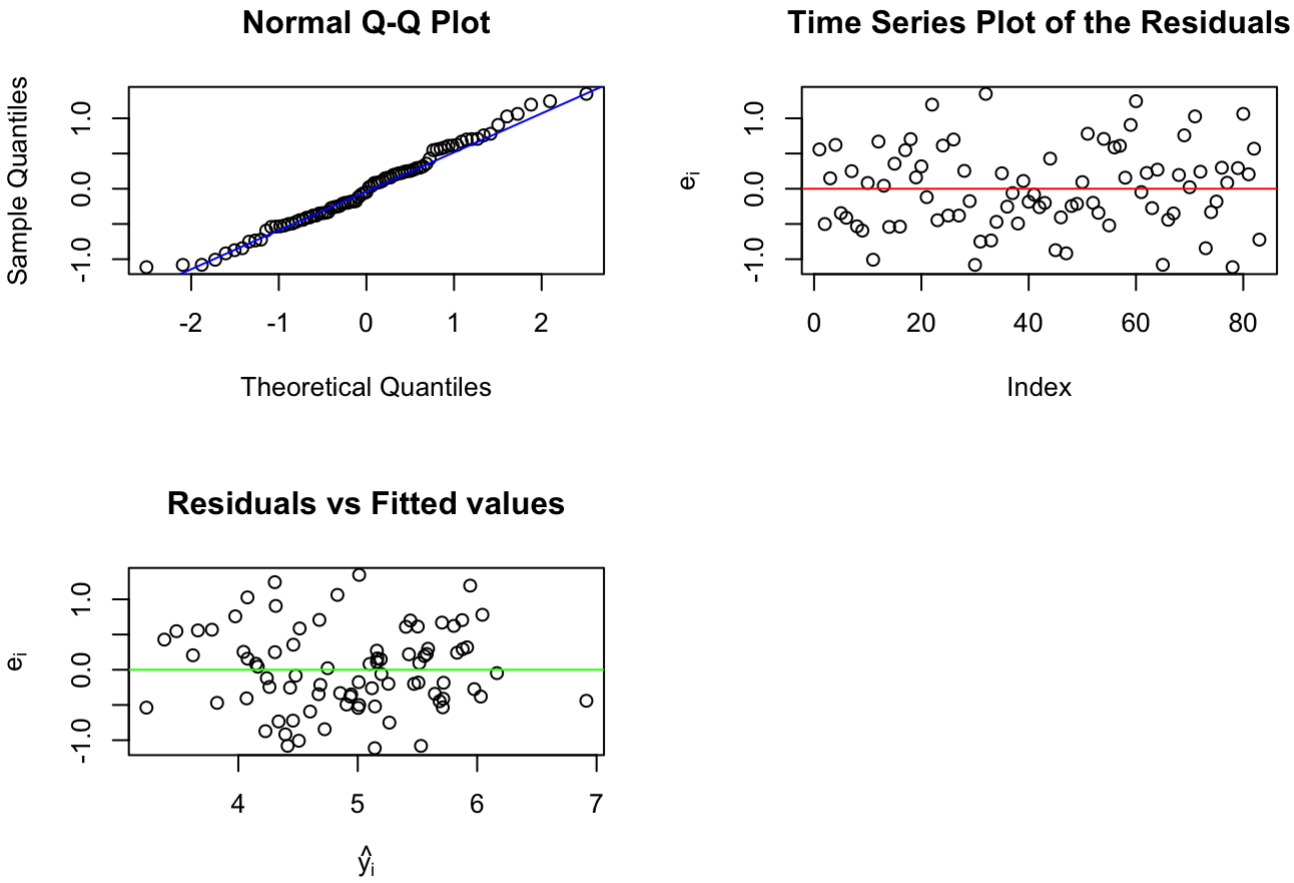
```
summary(best_m_2)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ veg_consumption + GDP + GHI +
##     unemployment + depression, data = data_reduced_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11348 -0.40917 -0.04789  0.33696  1.34689
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.903e+00  4.410e-01  15.652  < 2e-16 ***
## veg_consumption -1.957e-03  9.334e-04  -2.096  0.03933 *
## GDP              3.030e-05  1.115e-05   2.718  0.00811 **
## GHI             -4.988e-02  9.327e-03  -5.348 8.82e-07 ***
## unemployment    -3.119e-02  1.206e-02  -2.587  0.01158 *
## depression      -2.764e-01  1.081e-01  -2.556  0.01255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5943 on 77 degrees of freedom
## Multiple R-squared:  0.627,  Adjusted R-squared:  0.6027
## F-statistic: 25.88 on 5 and 77 DF,  p-value: 3.115e-15
```

It can be observed that the significance of the variables did not change, however coefficients did change which is due to the fact that observation 42 influenced the skewness of the model a lot. We also can observe that R_a^2 increased a little bit. Now let's proceed with verifying our 4 main assumptions.

Checking assumptions

```
par(mfrow=c(2,2))
qqnorm(best_m_2$residuals)
qqline(best_m_2$residuals, col = "blue")
plot(best_m_2$residuals,
     ylab=TeX("$e_i$"),
     main = "Time Series Plot of the Residuals")
abline(h=0, col="red")
plot(fitted(best_m_2),best_m_2$residuals,
     ylab=TeX("$e_i$"), xlab=TeX("\\hat{y_i}"),
     main = "Residuals vs Fitted values")
abline(h=0, col="green")
```



1. Normality The residuals must be normally distributed. It can be checked visually with the help of *Q-Q plot*. The Q-Q line is aproximately straight and diagonal, which suggests that our residuals are normally distributed.

2. Independence The residuals also must be independent of each other. We can check for independence by looking at the *Time-Series Plot of the Residuals*. The plot reveals no particular pattern, therefore we can conclude that out residuals are independent.

3. Linearity/Homogeneity of variance Our linear model should be a good fit. In order to check it we can refer to the *Resivuals vs Fitted values* plot. It also does not have any particular pattern and the points are evenly spread across 0 line, therefore we can conclude that the linearity assumption is satisfied. We also have to make sure that the variance of the residuals is hold constant, and we can check for it from the same plot. The plot seems to have some *megaphone* shape, which indicates that the variance might not be constant. Let's verify the homogeneity of variance with Breusch-Pagan Test.

```
bptest(best_m_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  best_m_2
## BP = 3.5299, df = 5, p-value = 0.6189
```

Since p-value is way larger than 0.05, the variance indeed must be non-constant. In order to eliminate this problem, we can use the Weighted-Least Squares. (WLS).

```
# Fitting the WLS
sd <- lm(abs(best_m_2$residuals) ~ veg_consumption+GDP+GHI+unemployment+depression, data
=data_reduced_5)
wls <- lm(life_satisfaction ~ veg_consumption+GDP+GHI+unemployment+depression, weights=1
/(abs(fitted(sd)))^2 , data=data_reduced_5)
```

Let's compare our previous model with WLS.

```
summary(best_m_2)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ veg_consumption + GDP + GHI +
##     unemployment + depression, data = data_reduced_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11348 -0.40917 -0.04789  0.33696  1.34689
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.903e+00  4.410e-01  15.652  < 2e-16 ***
## veg_consumption -1.957e-03  9.334e-04  -2.096  0.03933 *
## GDP              3.030e-05  1.115e-05   2.718  0.00811 **
## GHI             -4.988e-02  9.327e-03  -5.348 8.82e-07 ***
## unemployment    -3.119e-02  1.206e-02  -2.587  0.01158 *
## depression      -2.764e-01  1.081e-01  -2.556  0.01255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5943 on 77 degrees of freedom
## Multiple R-squared:  0.627,  Adjusted R-squared:  0.6027
## F-statistic: 25.88 on 5 and 77 DF,  p-value: 3.115e-15
```

```
summary(wls)
```

```
##
## Call:
## lm(formula = life_satisfaction ~ veg_consumption + GDP + GHI +
##     unemployment + depression, data = data_reduced_5, weights = 1/(abs(fitted(sd)))^
2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58140 -0.84154 -0.02773  0.72526  2.66640
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.126e+00  4.007e-01  17.782  < 2e-16 ***
## veg_consumption -2.346e-03  7.948e-04  -2.952  0.00418 **
## GDP              1.981e-05  8.052e-06   2.461  0.01610 *
## GHI             -5.674e-02  8.791e-03  -6.455 8.78e-09 ***
## unemployment    -3.532e-02  9.500e-03  -3.718  0.00038 ***
## depression      -2.567e-01  9.771e-02  -2.627  0.01038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.232 on 77 degrees of freedom
## Multiple R-squared:  0.6655, Adjusted R-squared:  0.6437
## F-statistic: 30.63 on 5 and 77 DF,  p-value: < 2.2e-16
```

From the above tables we can see that R_a^2 increased when we fitted the Weighted-Least Squares.

Discussion

The investigation started by conducting an exploratory data analysis. In this stage, it was revealed that some of the variables contained outliers and some of them were highly correlated with each other, which might cause multicollinearity. The initial model that we tried to fit was a simple additive model containing all 13 variables. In order to address the high correlation concern, the variance inflation factor (VIF) was calculated based on this model. The VIF scores were very high for some variables which suggests the presence of severe multicollinearity. In order to handle multicollinearity, 3 variables with the highest VIF scores were removed from the model one by one. The resultant model consisted of 10 variables and the VIF score for all of them was below 5. It is worth mentioning that the model with very small amount of multicollinearity is considered to be the one having all VIF scores to be close to 1. In our case, most of the variables have VIF scores near 1, but there are also scores in the range of 3 to 4.5, which is still a moderate amount of multicollinearity. It would have been a better choice to use ridge regression instead of dropping columns with the highest VIF scores, because even not severe, our model still contains variables causing moderate amount of multicollinearity which still might affect the performance of our model.

The variable selection process was then followed by the best subset method. Based on three metrics, namely R_a^2, Mallows's C_p and BIC, it was concluded that the best subset is a combination of `veg_consumption`, `GDP`, `GHI`, `unemployment`, and `depression`.

Following this, the model diagnostics were performed. Additive-Variable plots revealed that all 5 selected variables were significant to the model, and there was no need for higher-order terms. Interaction terms were also not needed because the plots "Residuals vs Interaction" did not show any non-linear pattern. We also investigated the outliers because some of them could influence the model a lot. The investigation was done with the help of Cook's distance chart and "Studentized Residuals vs leverage" plot. Both of the plots showed that observation 42 was an influential point. This point was further removed from the model such that the model could generalize to new observations well enough. The main 4 assumptions that the model had to follow also were checked. Normality, independence, and linearity revealed no problems, however, there seemed to be a problem with the homogeneity of variance. In order to solve this problem weighted least squares model was utilized.

Conclusion

The purpose of this project was to identify which factors affected the life satisfaction rate of people from developing countries in 2016. The investigation of this question was based on the data retrieved from Our World in Data (https://ourworldindata.org), which is a credible source. It is a scientific online publication that focuses on large global problems such as poverty, disease, inequality, etc. Its research team is based at the University of Oxford.

To conclude, from the above observations we can see that `veg_consumption`, `GDP`, `GHI`, `unemployment`, and `depression` in combination may explain the life expectancy rate very well. However, it should be pointed out that since this project is an observational study, it cannot prove causation. In order to make conclusions regarding causation, one should make research on the topic first and choose the most relevant variables. If I continued working with this project, I would definitely do that because in this way we could make decent conclusions from the study.