CMSE428 Lab Assignment 2

November 25, 2020

1 Data Science Assignment 2

2 Task 1

[1]:
```
%matplotlib inline

import pandas as pd
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from heatmap import heatmap, corrplot
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from scipy import stats
import statsmodels.api as sm
%config InlineBackend.figure_format = 'retina'
pd.set_option('display.float_format', lambda x: '%.5f' % x)
```

2.1 Dataset Loaded in

[2]:
```
dataset = pd.read_csv('Life Expectancy Data 1.csv')
```

2.2 Original Dataset

[3]:
```
dataset.head()
```

[3]:

	Country	Year	Status	Life Expectancy	Adult Mortality	\
0	Afghanistan	2015	Developing	65.00000	263.00000	
1	Afghanistan	2014	Developing	59.90000	271.00000	
2	Afghanistan	2013	Developing	59.90000	268.00000	
3	Afghanistan	2012	Developing	59.50000	272.00000	
4	Afghanistan	2011	Developing	59.20000	275.00000	

	Infant Deaths	Alcohol	Percentage Expenditure	Hepatitis B	Measles	…	\
0	62	0.01000	71.27962	65.00000	1154	…	
1	64	0.01000	73.52358	62.00000	492	…	

```
2             66  0.01000               73.21924     64.00000      430  …
3             69  0.01000               78.18422     67.00000     2787  …
4             71  0.01000                7.09711     68.00000     3013  …

     Polio  Total Expenditure  Diphtheria  HIV/AIDS       GDP     Population  \
0  6.00000            8.16000    65.00000   0.10000 584.25921 33736494.00000
1 58.00000            8.18000    62.00000   0.10000 612.69651   327582.00000
2 62.00000            8.13000    64.00000   0.10000 631.74498 31731688.00000
3 67.00000            8.52000    67.00000   0.10000 669.95900  3696958.00000
4 68.00000            7.87000    68.00000   0.10000  63.53723  2978599.00000

   Thinness 1-19 Years  Thinness 5-9 Years  Income Composition of Resources  \
0              17.20000             17.30000                          0.47900
1              17.50000             17.50000                          0.47600
2              17.70000             17.70000                          0.47000
3              17.90000             18.00000                          0.46300
4              18.20000             18.20000                          0.45400

   Schooling
0   10.10000
1   10.00000
2    9.90000
3    9.80000
4    9.50000

[5 rows x 22 columns]
```

2.3 a) Preprocess the dataset to impute missing values.

2.4 Checking which columns have missing Data

By running the below code, we can see the number of missing data we have.

[4]:
```
dataset.isnull().sum()
```

[4]:
```
Country                             0
Year                                0
Status                              0
Life Expectancy                    10
Adult Mortality                    10
Infant Deaths                       0
Alcohol                           194
Percentage Expenditure              0
Hepatitis B                       553
Measles                             0
BMI                                34
Under-Five Deaths                   0
Polio                              19
```

```
Total Expenditure                  226
Diphtheria                          19
HIV/AIDS                             0
GDP                                448
Population                         652
Thinness 1-19 Years                 34
Thinness 5-9 Years                  34
Income Composition of Resources    167
Schooling                          163
dtype: int64
```

2.5 Seperating Columns to be filled by Most Frequent and Mean

I have used Most Frequent method to fill the missing INTEGER data, to avoid stuff like "20.5 Deaths". I have used Mean method to fill the rest of the missing data, with decimal points.

[5]:
```
mean_columns = dataset[['Life Expectancy','Alcohol','BMI','Total␣
↪Expenditure','GDP','Thinness 1-19 Years','Thinness 5-9 Years','Income␣
↪Composition of Resources','Schooling']]
mfrequent_columns = dataset[['Adult Mortality','Hepatitis␣
↪B','Polio','Diphtheria','Population']]

for column in mean_columns:
    dataset[column].fillna(dataset[column].mean(), inplace=True)

for column in mfrequent_columns:
    dataset[column].fillna(dataset[column].value_counts().index[0],␣
↪inplace=True)
```

2.5.1 By running this code, we can verify that we filled in the missing data successfully.

[6]:
```
dataset.isnull().sum()
```

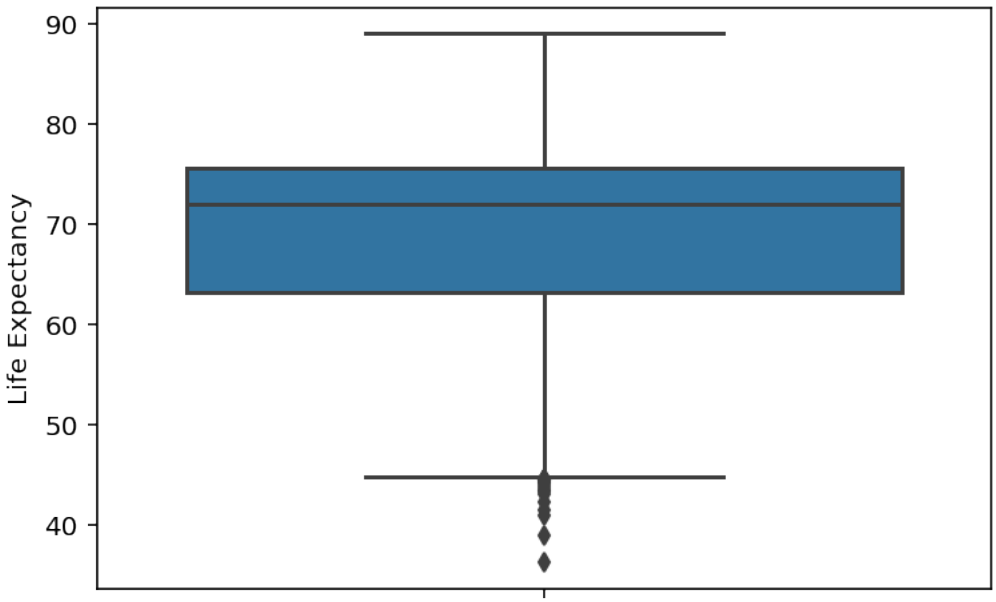
[6]:
```
Country                          0
Year                             0
Status                           0
Life Expectancy                  0
Adult Mortality                  0
Infant Deaths                    0
Alcohol                          0
Percentage Expenditure           0
Hepatitis B                      0
Measles                          0
BMI                              0
Under-Five Deaths                0
Polio                            0
```

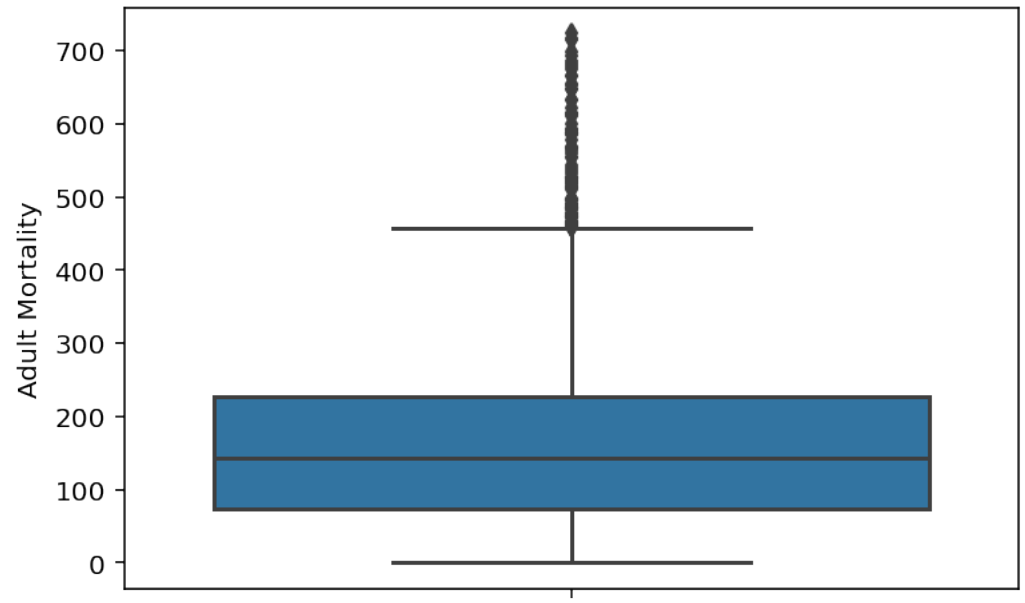
```
Total Expenditure                  0
Diphtheria                         0
HIV/AIDS                           0
GDP                                0
Population                         0
Thinness 1-19 Years                0
Thinness 5-9 Years                 0
Income Composition of Resources    0
Schooling                          0
dtype: int64
```

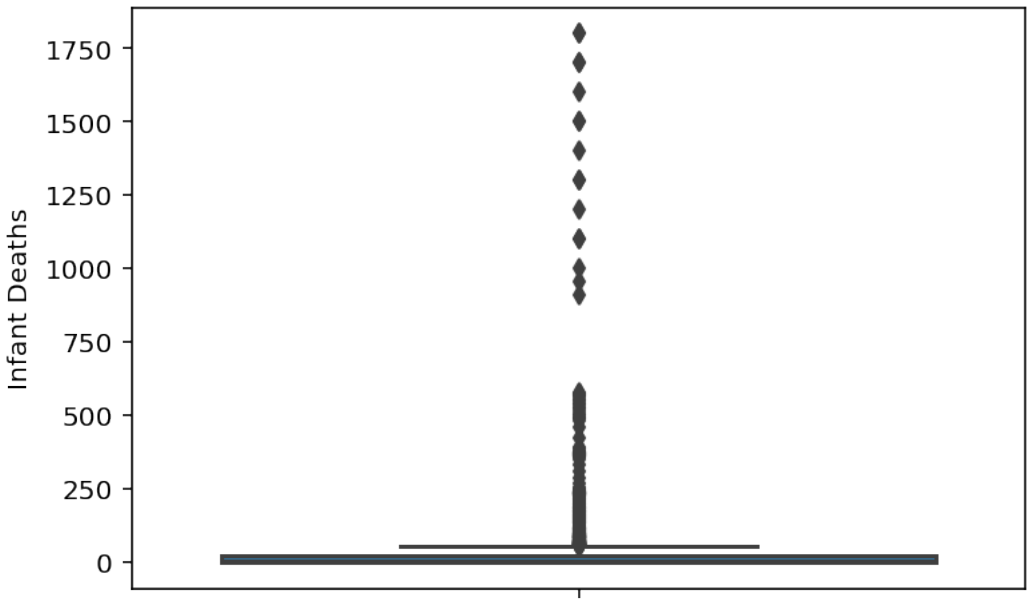
2.6 b) Use boxplots to check for outliers.

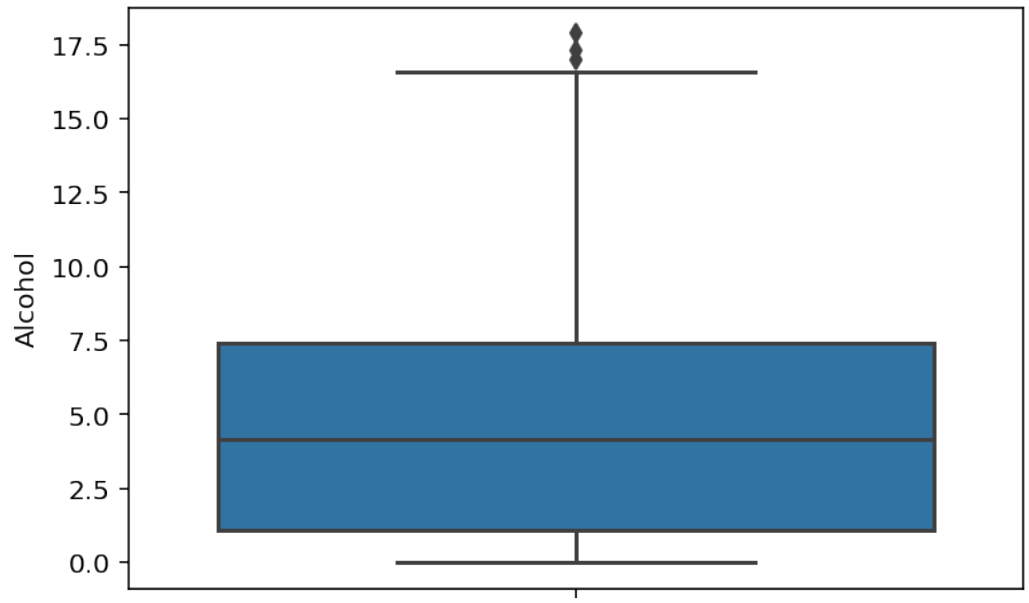
Using the boxplot function of the seaborn library, I was able to show the box plot and outliers of each column.

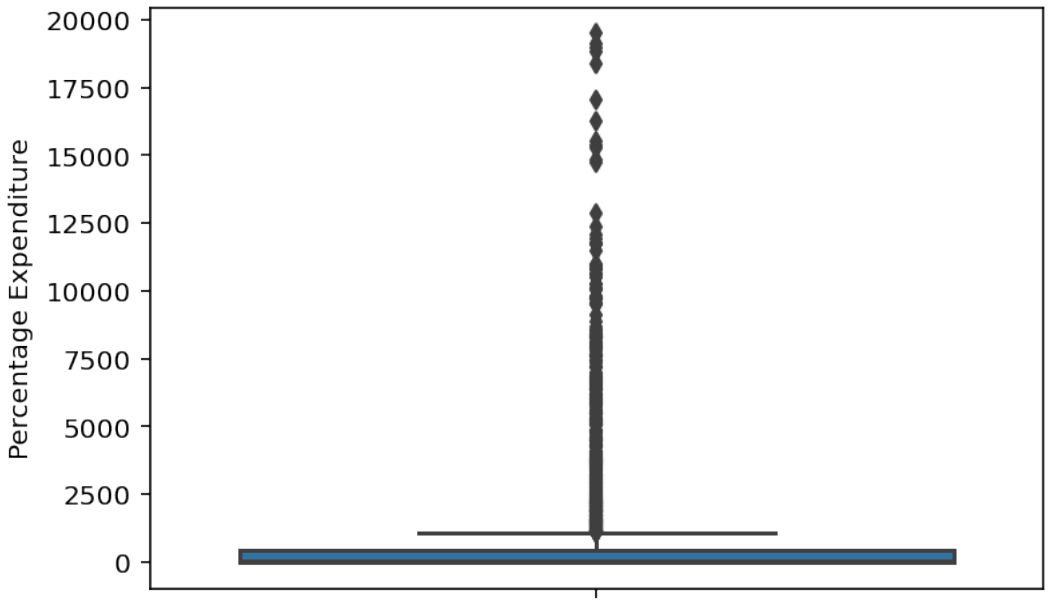
[7]:
```
columns = ['Life Expectancy','Adult Mortality','Infant Deaths','Alcohol',␣
 ↪'Percentage Expenditure','Hepatitis B','Measles','BMI', 'Under-Five Deaths',␣
 ↪'Polio', 'Total Expenditure', 'Diphtheria', 'HIV/AIDS',␣
 ↪'GDP','Population','Thinness 1-19 Years', 'Thinness 5-9 Years', 'Income␣
 ↪Composition of Resources', 'Schooling']
for column in columns:
    sns.boxplot(data=dataset, y=column)
    plt.show()
```

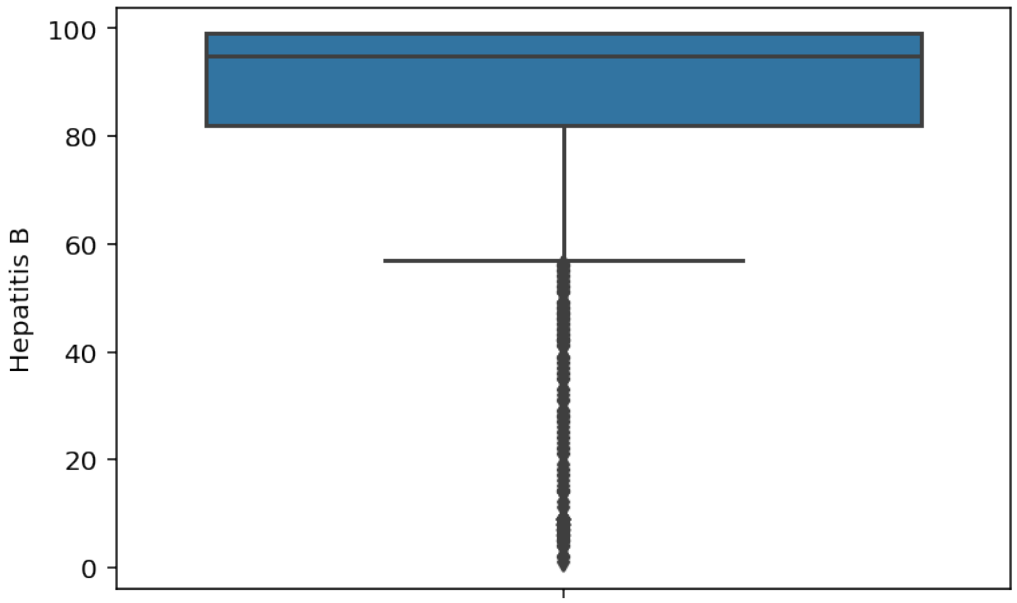


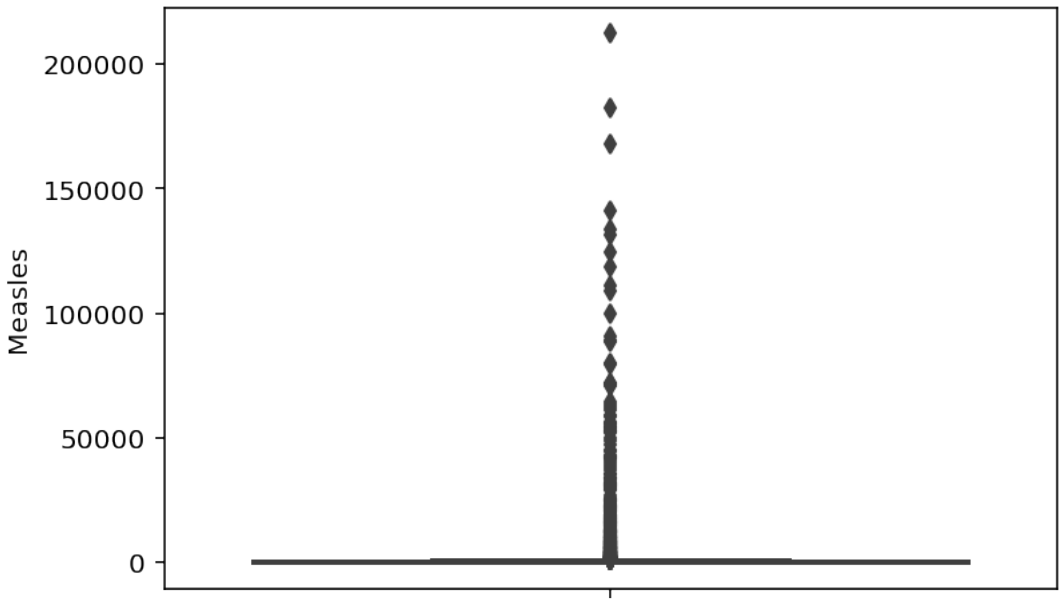

Adult Mortality
700
600
500
400
300
200
100
0

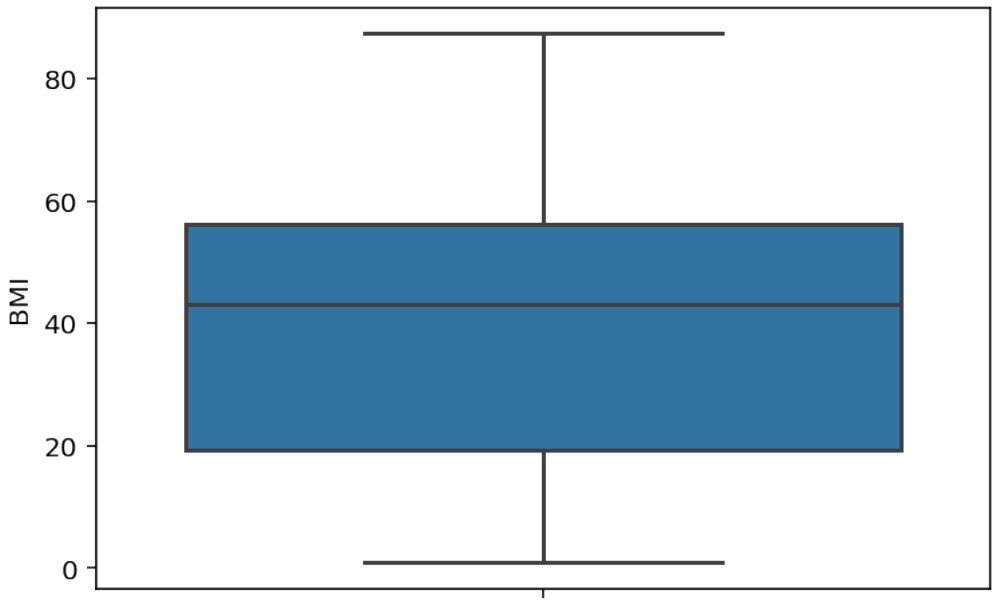

Infant Deaths
1750
1500
1250
1000
750
500
250
0

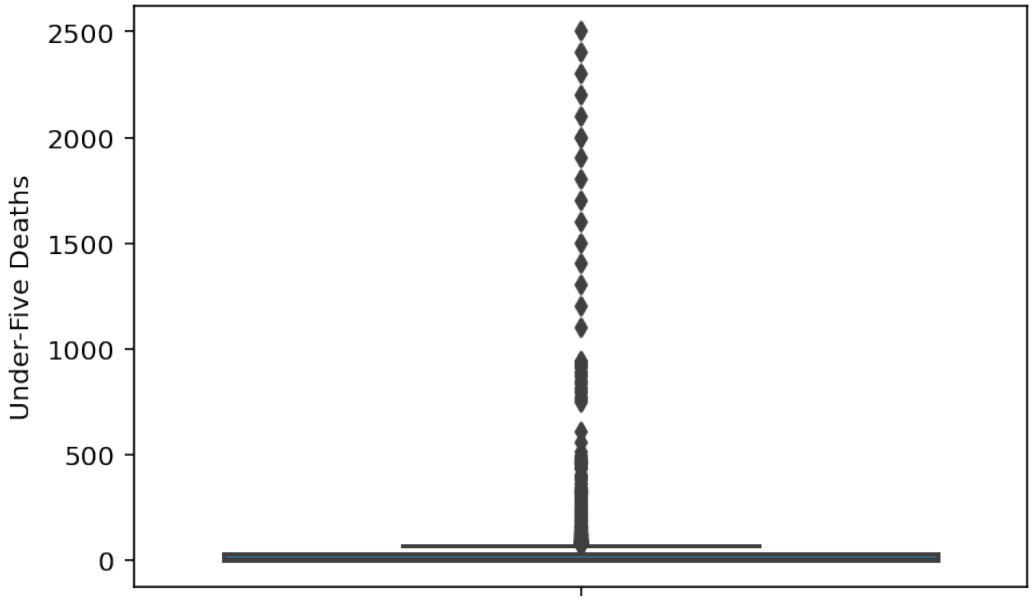

Alcohol
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

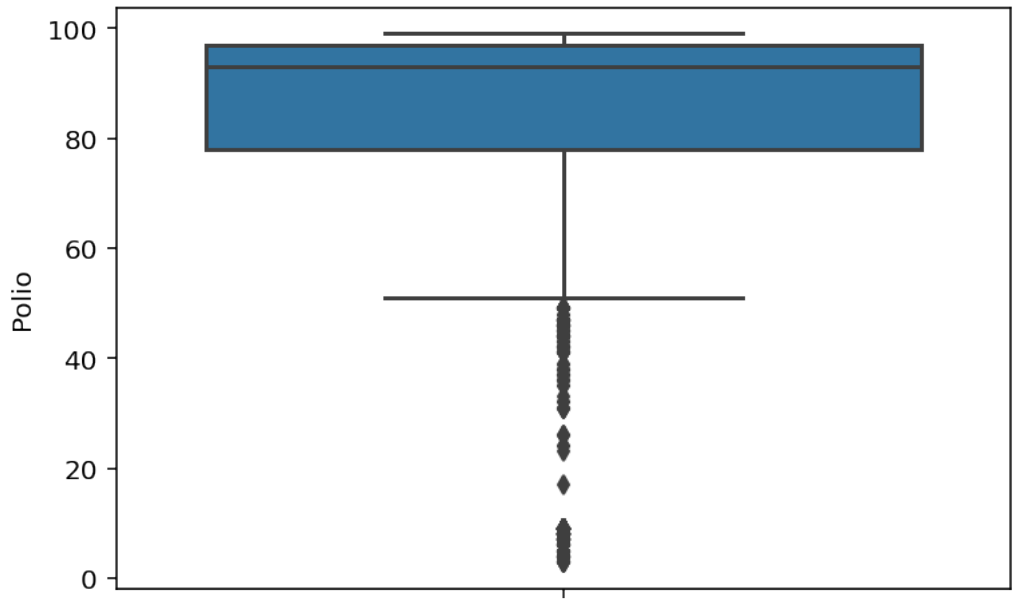

Percentage Expenditure
20000
17500
15000
12500
10000
7500
5000
2500
0

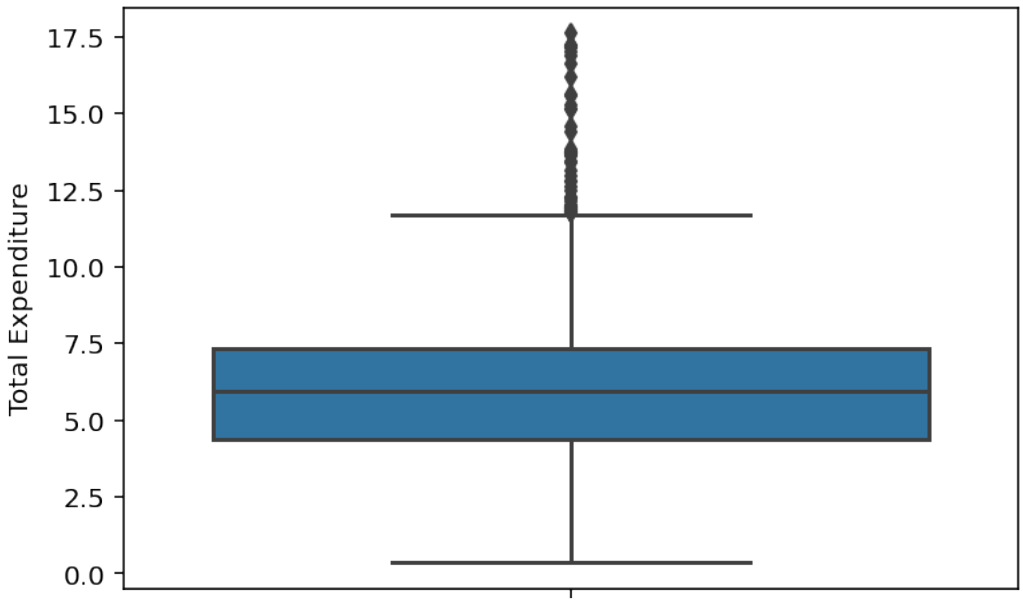

Hepatitis B
100
80
60
40
20
0

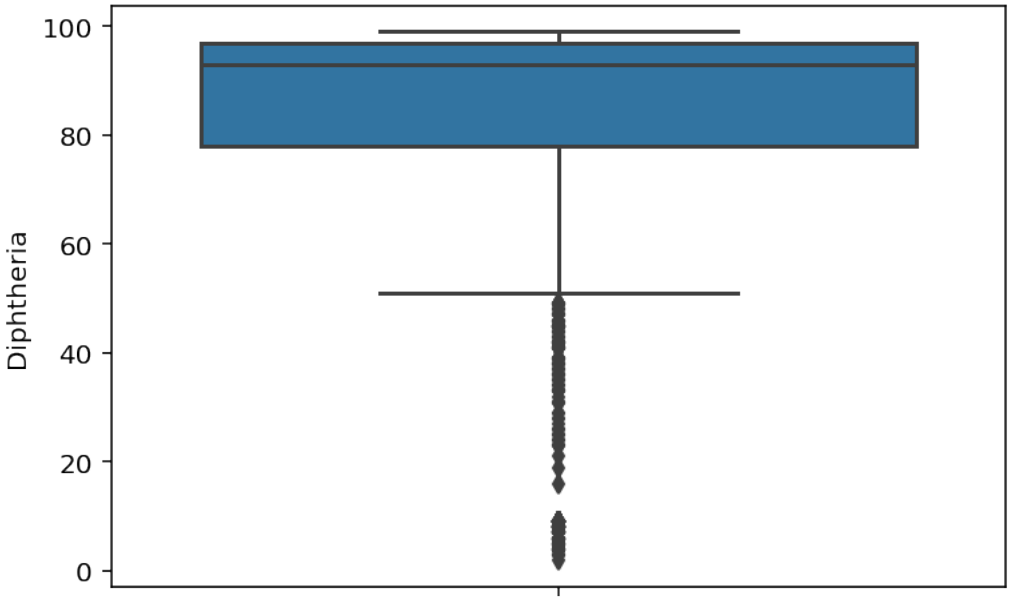

Measles
200000
150000
100000
50000
0

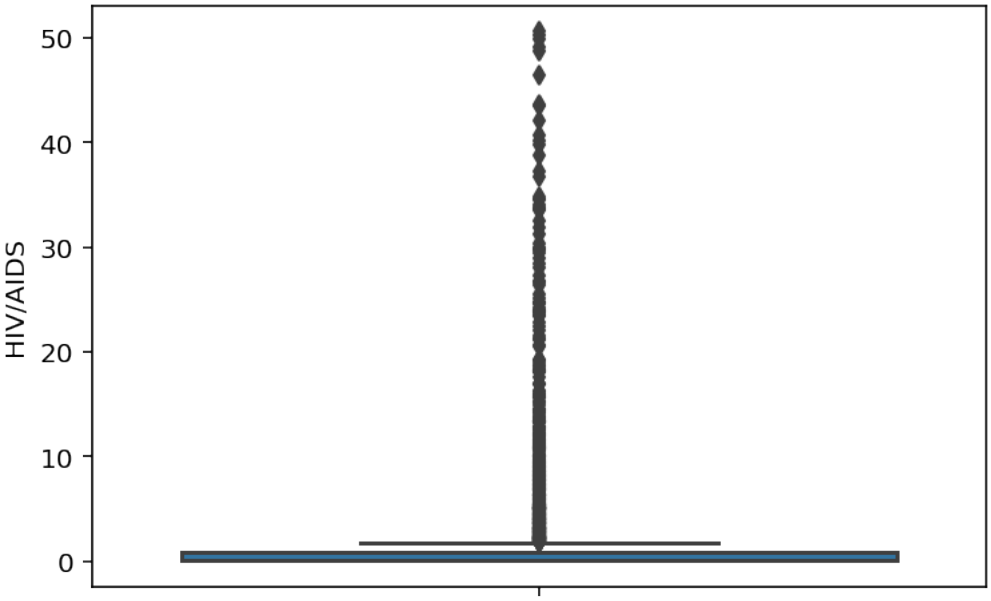

BMI
80
60
40
20
0

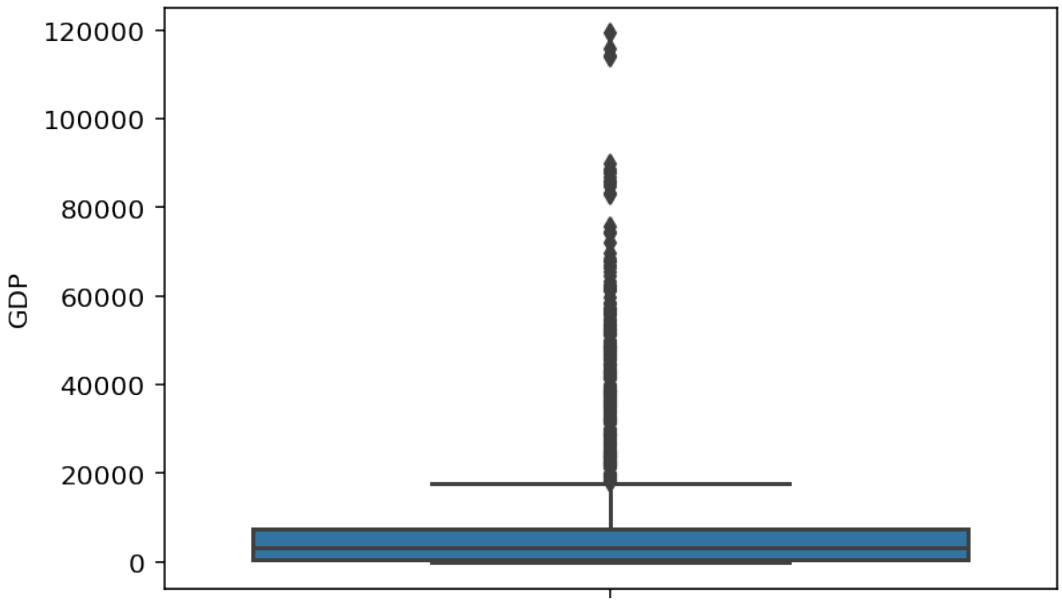

Under-Five Deaths
2500
2000
1500
1000
500
0

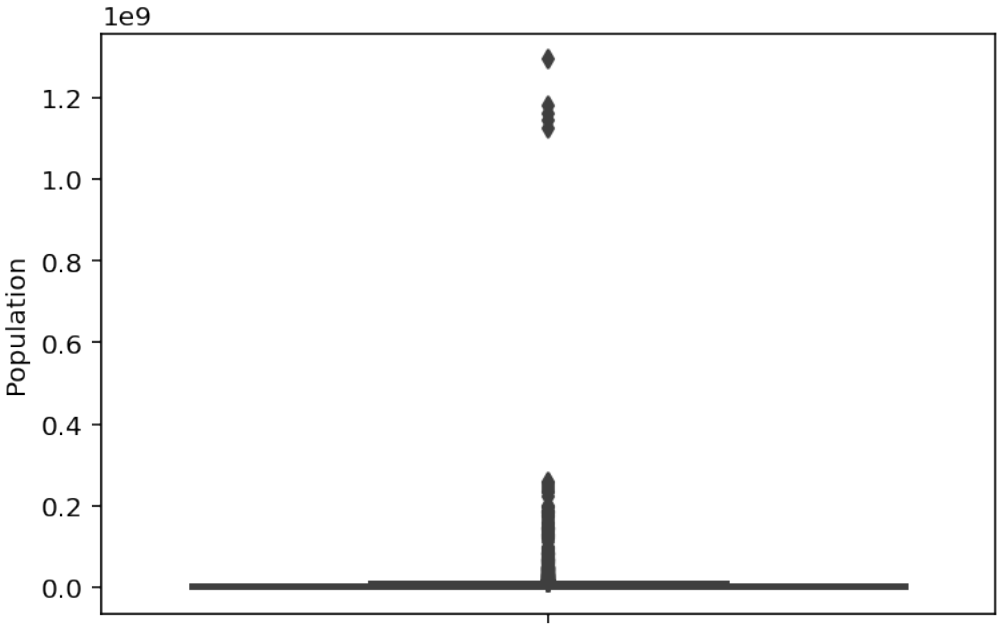

Polio
100
80
60
40
20
0

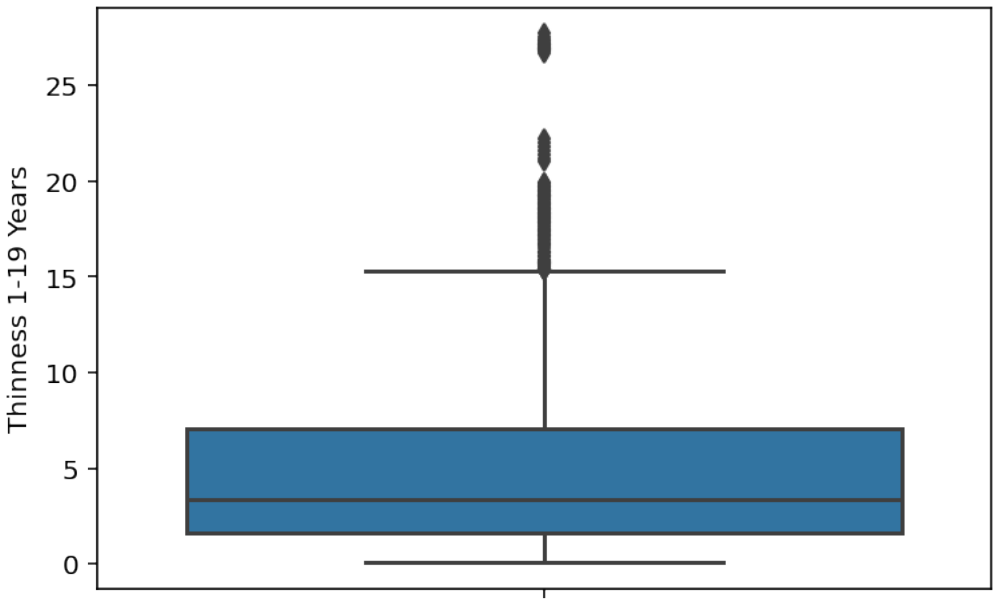

Total Expenditure
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

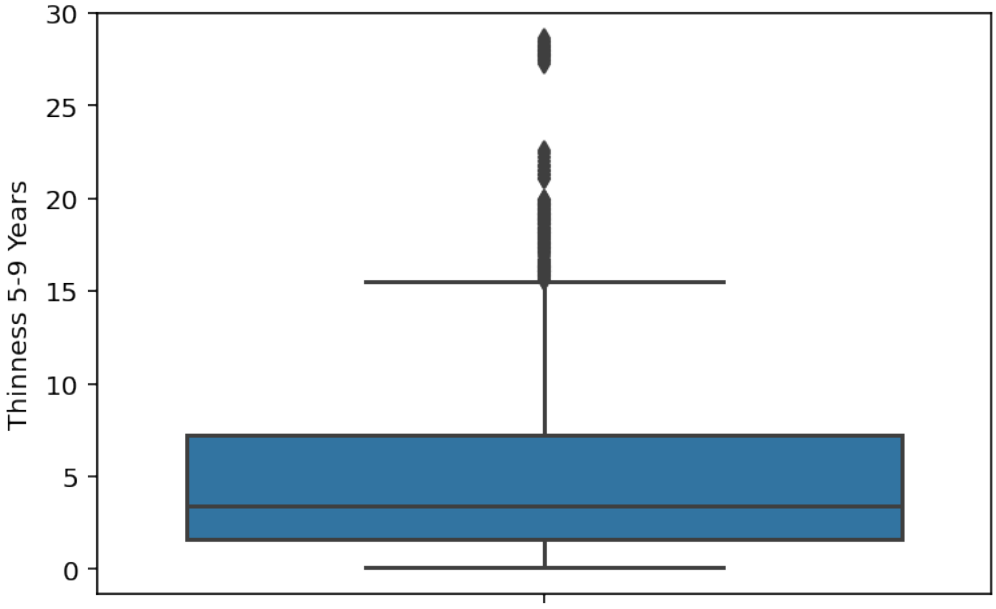

Diphtheria
100
80
60
40
20
0

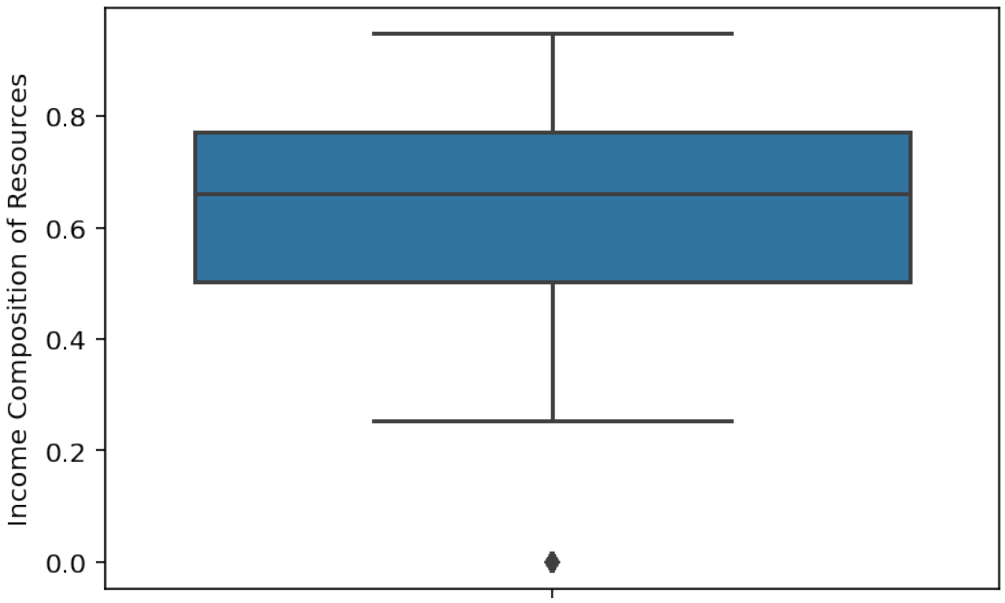

HIV/AIDS
50
40
30
20
10
0

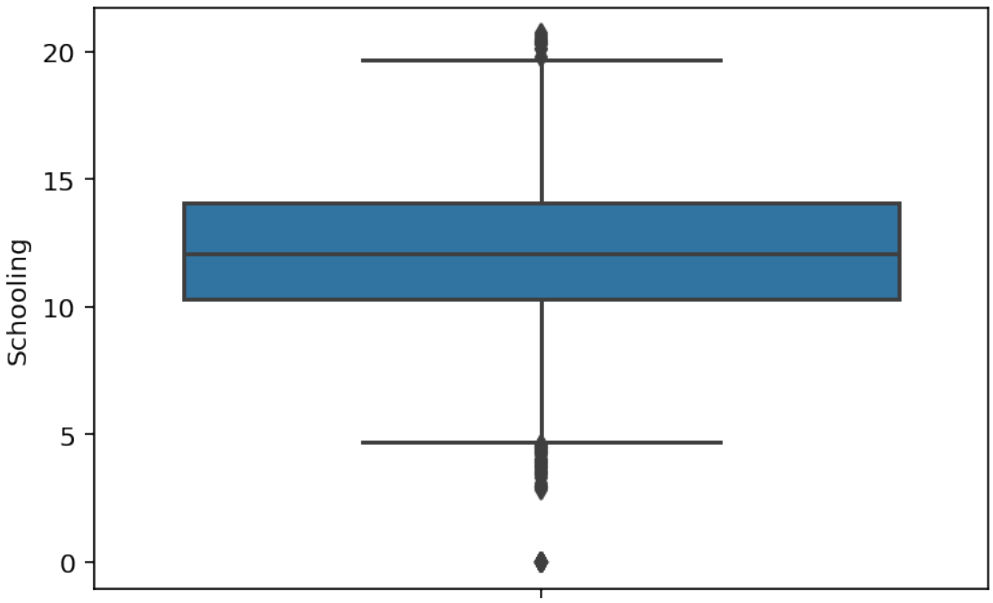

GDP
120000
100000
80000
60000
40000
20000
0


1e9
Population
1.2
1.0
0.8
0.6
0.4
0.2
0.0


Thinness 1-19 Years
25
20
15
10
5
0


Thinness 5-9 Years
30
25
20
15
10
5
0


Income Composition of Resources
0.0
0.2
0.4
0.6
0.8


Schooling
0
5
10
15
20

3 c) Provide a correlation plot for the variables in the dataset, including the dependent variable. Exclude the columns A, B and C. Comment on the strengths on the predictors. Comment on the correlations between predictors.

Checking the correlations between the predictors, of course they correlate best with themselves. The more blue and bigger the squares are, the better they correlate which each other.

For example if we look at Under-Five Deaths and Infant deaths, they have high positive correlation with each other. Schooling and Income Composition of Resources have high positive correlation with each other. Thinness 5-9 Years and Thinness 1-19 Years have high positive correlation with each other.

For example if we look at BMI and Thinness 5-9 Years and Thinness 1-19 Years, they have high negative correlation with each other. Life expectancy and Adult Mortality have high negative correlation with each other.

[8]:
```
plt.figure(figsize=(8, 8))
corrplot(dataset.corr());
```

4 Task 2

4.1 a) Determine which independent variable is the most related with the dependent variable by developing a linear regression model between the dependent variable and each independent variable.

To determine which independent variable is the most related, we first have to create the models and then calculated squared_adj for each of them.

Declared the variables and seperated the predictors.

[9]:
```
model_scores = {}
predictors = ['Adult Mortality','Infant Deaths','Alcohol', 'Percentage␣
↪Expenditure','Hepatitis B','Measles','BMI', 'Under-Five Deaths', 'Polio',␣
↪'Total Expenditure', 'Diphtheria', 'HIV/AIDS', 'GDP','Population','Thinness␣
↪1-19 Years', 'Thinness 5-9 Years', 'Income Composition of Resources',␣
↪'Schooling']
```

In this loop, each predictor in the predictors list gets into Linear Regression and then statsmodels module helps us calculate adjusted rsquared and then we add the scores into the model_scores dictionary to later be shown as a table.

[10]:
```
for predictor in predictors:
    x=dataset[[predictor]]
    y=dataset[['Life Expectancy']]

    X = sm.add_constant(x)
    reg = sm.OLS(y, X).fit()

    model_scores[x.columns[0]] = reg.rsquared_adj
```

4.2 b) Rank the predictors according to the Adjusted R2 values of the linear models obtained.

Now we put the scores into the DataFrame to display. Higher Adjusted R2 Score means better. Also to answer the previous question, higher the value is on the list, the better related it is.

[11]:
```
r2_scores_df = pd.DataFrame.from_dict(model_scores, orient='index',␣
↪columns=['Adjusted R2 Scores'])
r2_scores_df.sort_values(by='Adjusted R2 Scores',ascending=False)
```

[11]:

	Adjusted R2 Scores
Schooling	0.51115
Adult Mortality	0.48226
Income Composition of Resources	0.47936
BMI	0.31253

```
HIV/AIDS                          0.30941
Thinness 1-19 Years               0.22267
Diphtheria                        0.22083
Thinness 5-9 Years                0.21748
Polio                             0.20802
GDP                               0.18505
Alcohol                           0.15306
Percentage Expenditure            0.14547
Under-Five Deaths                 0.04918
Total Expenditure                 0.04293
Infant Deaths                     0.03830
Measles                           0.02450
Hepatitis B                       0.02144
Population                        0.00056
```

5 Task 3

5.1 a) Compute a linear regression model using all dependent variables and report the Adjusted R2 value.

Here we have to use Multiple Linear Regression technique, luckily statsmodels support multiple linear regression. From there we can calculate the rsquared_adj as shown at the output.

[12]:
```
x=dataset[predictors]
y=dataset[['Life Expectancy']]

X = sm.add_constant(x)
all_reg = sm.OLS(y, X).fit()

all_reg.rsquared_adj
```

[12]:
```
0.8165300412769076
```

5.2 b) Display the model (i.e. coefficients) and comment about the relative importance of the predictors by considering the p-values of the predictors

By using the summary() function of statsmodels, we can check the stats of the linear regression. If we take a look at "P>|t|" column of the output, we can see the p-values of the predictors.

[13]:
```
all_reg.summary()
```

[13]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life Expectancy   R-squared:                       0.818
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     727.2
```

```
Date:                Wed, 25 Nov 2020   Prob (F-statistic):               0.00
Time:                        23:07:08   Log-Likelihood:                -8285.0
No. Observations:                2938   AIC:                         1.661e+04
Df Residuals:                    2919   BIC:                         1.672e+04
Df Model:                          18
Covariance Type:            nonrobust
================================================================================
                                 coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
const                          54.5184      0.583     93.571      0.000
53.376      55.661
Adult Mortality                -0.0202      0.001    -25.477      0.000
-0.022      -0.019
Infant Deaths                   0.0988      0.008     11.650      0.000
0.082       0.115
Alcohol                         0.1268      0.024      5.263      0.000
0.080       0.174
Percentage Expenditure          0.0001   8.43e-05      1.769      0.077
-1.62e-05       0.000
Hepatitis B                    -0.0143      0.004     -4.019      0.000
-0.021      -0.007
Measles                     -1.987e-05   7.69e-06     -2.583      0.010
-3.5e-05   -4.79e-06
BMI                             0.0418      0.005      8.433      0.000
0.032       0.052
Under-Five Deaths              -0.0738      0.006    -11.876      0.000
-0.086      -0.062
Polio                           0.0287      0.004      6.426      0.000
0.020       0.037
Total Expenditure               0.0845      0.034      2.490      0.013
0.018       0.151
Diphtheria                      0.0394      0.005      8.514      0.000
0.030       0.048
HIV/AIDS                       -0.4723      0.018    -26.792      0.000
-0.507      -0.438
GDP                          3.573e-05    1.3e-05      2.744      0.006
1.02e-05    6.13e-05
Population                  -1.422e-10    1.7e-09     -0.084      0.933
-3.47e-09    3.19e-09
Thinness 1-19 Years            -0.0807      0.051     -1.595      0.111
-0.180       0.019
Thinness 5-9 Years              0.0003      0.050      0.006      0.995
-0.098       0.098
Income Composition of Resources     5.9024      0.638      9.256      0.000
```

```
4.652       7.153
Schooling                          0.6843      0.042     16.373      0.000
0.602       0.766
==============================================================================
Omnibus:                      131.523   Durbin-Watson:                   0.725
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              377.435
Skew:                          -0.171   Prob(JB):                     1.10e-82
Kurtosis:                       4.722   Cond. No.                     4.84e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 4.84e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

5.3 c) Using either AIC or p-values, discard the weakly related variables using either forward or backward selection

By using this script I have found online, I could implement forward and backward selection into Python.

[14]:
```
#Copyright 2019 Sinan Talha Hascelik
#
#Licensed under the Apache License, Version 2.0 (the "License");
#you may not use this file except in compliance with the License.
#You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
#Unless required by applicable law or agreed to in writing, software
#distributed under the License is distributed on an "AS IS" BASIS,
#WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
#See the License for the specific language governing permissions and
#limitations under the License.

import numpy as np
import pandas as pd
import statsmodels.formula.api as sm
import statsmodels.api as sm

def forwardSelection(X, y, model_type ="linear",elimination_criteria = "aic",␣
↪varchar_process = "dummy_dropfirst", sl=0.05):
    """
    Forward Selection is a function, based on regression models, that returns␣
↪significant features and selection iterations.\n
```

```
    Required Libraries: pandas, numpy, statmodels

    Parameters
    ----------
    X : Independent variables (Pandas Dataframe)\n
    y : Dependent variable (Pandas Series, Pandas Dataframe)\n
    model_type : 'linear' or 'logistic'\n
    elimination_criteria : 'aic', 'bic', 'r2', 'adjr2' or None\n
        'aic' refers Akaike information criterion\n
        'bic' refers Bayesian information criterion\n
        'r2' refers R-squared (Only works on linear model type)\n
        'r2' refers Adjusted R-squared (Only works on linear model type)\n
    varchar_process : 'drop', 'dummy' or 'dummy_dropfirst'\n
        'drop' drops varchar features\n
        'dummy' creates dummies for all levels of all varchars\n
        'dummy_dropfirst' creates dummies for all levels of all varchars, and␣
↪drops first levels\n
    sl : Significance Level (default: 0.05)\n


    Returns
    -------
    columns(list), iteration_logs(str)\n\n
    Not Returns a Model


    Tested On
    ---------
    Python v3.6.7, Pandas v0.23.4, Numpy v1.15.04, StatModels v0.9.0


    See Also
    --------
    https://en.wikipedia.org/wiki/Stepwise_regression
    """
    X = __varcharProcessing__(X,varchar_process = varchar_process)
    return __forwardSelectionRaw__(X, y, model_type =␣
↪model_type,elimination_criteria = elimination_criteria , sl=sl)

def backwardSelection(X, y, model_type ="linear",elimination_criteria = "aic",␣
↪varchar_process = "dummy_dropfirst", sl=0.05):
    """
    Backward Selection is a function, based on regression models, that returns␣
↪significant features and selection iterations.\n
    Required Libraries: pandas, numpy, statmodels

    Parameters
```

```
    ----------
    X : Independent variables (Pandas Dataframe)\n
    y : Dependent variable (Pandas Series, Pandas Dataframe)\n
    model_type : 'linear' or 'logistic'\n
    elimination_criteria : 'aic', 'bic', 'r2', 'adjr2' or None\n
        'aic' refers Akaike information criterion\n
        'bic' refers Bayesian information criterion\n
        'r2' refers R-squared (Only works on linear model type)\n
        'r2' refers Adjusted R-squared (Only works on linear model type)\n
    varchar_process : 'drop', 'dummy' or 'dummy_dropfirst'\n
        'drop' drops varchar features\n
        'dummy' creates dummies for all levels of all varchars\n
        'dummy_dropfirst' creates dummies for all levels of all varchars, and␣
↪drops first levels\n
    sl : Significance Level (default: 0.05)\n


    Returns
    -------
    columns(list), iteration_logs(str)\n\n
    Not Returns a Model


    Tested On
    ---------
    Python v3.6.7, Pandas v0.23.4, Numpy v1.15.04, StatModels v0.9.0


    See Also
    --------
    https://en.wikipedia.org/wiki/Stepwise_regression
    """
    X = __varcharProcessing__(X,varchar_process = varchar_process)
    return __backwardSelectionRaw__(X, y, model_type =␣
↪model_type,elimination_criteria = elimination_criteria , sl=sl)

def __varcharProcessing__(X, varchar_process = "dummy_dropfirst"):

    dtypes = X.dtypes
    if varchar_process == "drop":
        X = X.drop(columns = dtypes[dtypes == np.object].index.tolist())
        print("Character Variables (Dropped):", dtypes[dtypes == np.object].
↪index.tolist())
    elif varchar_process == "dummy":
        X = pd.get_dummies(X,drop_first=False)
        print("Character Variables (Dummies Generated):", dtypes[dtypes == np.
↪object].index.tolist())
```

```
    elif varchar_process == "dummy_dropfirst":
        X = pd.get_dummies(X,drop_first=True)
        print("Character Variables (Dummies Generated, First Dummies Dropped):
 ↪", dtypes[dtypes == np.object].index.tolist())
    else:
        X = pd.get_dummies(X,drop_first=True)
        print("Character Variables (Dummies Generated, First Dummies Dropped):
 ↪", dtypes[dtypes == np.object].index.tolist())

    X["intercept"] = 1
    cols = X.columns.tolist()
    cols = cols[-1:] + cols[:-1]
    X = X[cols]

    return X

def __forwardSelectionRaw__(X, y, model_type ="linear",elimination_criteria =␣
 ↪"aic", sl=0.05):

    iterations_log = ""
    cols = X.columns.tolist()

    def regressor(y,X, model_type=model_type):
        if model_type == "linear":
            regressor = sm.OLS(y, X).fit()
        elif model_type == "logistic":
            regressor = sm.Logit(y, X).fit()
        else:
            print("\nWrong Model Type : "+ model_type +"\nLinear model type is␣
 ↪seleted.")
            model_type = "linear"
            regressor = sm.OLS(y, X).fit()
        return regressor

    selected_cols = ["intercept"]
    other_cols = cols.copy()
    other_cols.remove("intercept")

    model = regressor(y, X[selected_cols])

    if elimination_criteria == "aic":
        criteria = model.aic
    elif elimination_criteria == "bic":
        criteria = model.bic
    elif elimination_criteria == "r2" and model_type =="linear":
        criteria = model.rsquared
    elif elimination_criteria == "adjr2" and model_type =="linear":
```

```
        criteria = model.rsquared_adj


    for i in range(X.shape[1]):
        pvals = pd.DataFrame(columns = ["Cols","Pval"])
        for j in other_cols:
            model = regressor(y, X[selected_cols+[j]])
            pvals = pvals.append(pd.DataFrame([[j, model.pvalues[j]]],columns =␣
↪["Cols","Pval"]),ignore_index=True)
        pvals = pvals.sort_values(by = ["Pval"]).reset_index(drop=True)
        pvals = pvals[pvals.Pval<=sl]
        if pvals.shape[0] > 0:

            model = regressor(y, X[selected_cols+[pvals["Cols"][0]]])
            iterations_log += str("\nEntered : "+pvals["Cols"][0] + "\n")
            iterations_log += "\n\n"+str(model.summary())+"\nAIC: "+ str(model.
↪aic) + "\nBIC: "+ str(model.bic)+"\n\n"


            if  elimination_criteria == "aic":
                new_criteria = model.aic
                if new_criteria < criteria:
                    print("Entered :", pvals["Cols"][0], "\tAIC :", model.aic)
                    selected_cols.append(pvals["Cols"][0])
                    other_cols.remove(pvals["Cols"][0])
                    criteria = new_criteria
                else:
                    print("break : Criteria")
                    break
            elif  elimination_criteria == "bic":
                new_criteria = model.bic
                if new_criteria < criteria:
                    print("Entered :", pvals["Cols"][0], "\tBIC :", model.bic)
                    selected_cols.append(pvals["Cols"][0])
                    other_cols.remove(pvals["Cols"][0])
                    criteria = new_criteria
                else:
                    print("break : Criteria")
                    break
            elif  elimination_criteria == "r2" and model_type =="linear":
                new_criteria = model.rsquared
                if new_criteria > criteria:
                    print("Entered :", pvals["Cols"][0], "\tR2 :", model.
↪rsquared)
                    selected_cols.append(pvals["Cols"][0])
                    other_cols.remove(pvals["Cols"][0])
                    criteria = new_criteria
```

```
                else:
                    print("break : Criteria")
                    break
            elif  elimination_criteria == "adjr2" and model_type =="linear":
                new_criteria = model.rsquared_adj
                if new_criteria > criteria:
                    print("Entered :", pvals["Cols"][0], "\tAdjR2 :", model.
↪rsquared_adj)
                    selected_cols.append(pvals["Cols"][0])
                    other_cols.remove(pvals["Cols"][0])
                    criteria = new_criteria
                else:
                    print("Break : Criteria")
                    break
            else:
                print("Entered :", pvals["Cols"][0])
                selected_cols.append(pvals["Cols"][0])
                other_cols.remove(pvals["Cols"][0])

        else:
            print("Break : Significance Level")
            break

    model = regressor(y, X[selected_cols])
    if elimination_criteria == "aic":
        criteria = model.aic
    elif elimination_criteria == "bic":
        criteria = model.bic
    elif elimination_criteria == "r2" and model_type =="linear":
        criteria = model.rsquared
    elif elimination_criteria == "adjr2" and model_type =="linear":
        criteria = model.rsquared_adj

    print(model.summary())
    print("AIC: "+str(model.aic))
    print("BIC: "+str(model.bic))
    print("Final Variables:", selected_cols)

    return selected_cols, iterations_log

def __backwardSelectionRaw__(X, y, model_type ="linear",elimination_criteria =␣
↪"aic", sl=0.05):

    iterations_log = ""
    last_eleminated = ""
    cols = X.columns.tolist()
```

```
    def regressor(y,X, model_type=model_type):
        if model_type =="linear":
            regressor = sm.OLS(y, X).fit()
        elif model_type == "logistic":
            regressor = sm.Logit(y, X).fit()
        else:
            print("\nWrong Model Type : "+ model_type +"\nLinear model type is␣
↪seleted.")
            model_type = "linear"
            regressor = sm.OLS(y, X).fit()
        return regressor
    for i in range(X.shape[1]):
        if i != 0 :
            if elimination_criteria == "aic":
                criteria = model.aic
                new_model = regressor(y,X)
                new_criteria = new_model.aic
                if criteria < new_criteria:
                    print("Regained : ", last_eleminated)
                    iterations_log += "\n"+str(new_model.summary())+"\nAIC: "+␣
↪str(new_model.aic) + "\nBIC: "+ str(new_model.bic)+"\n"
                    iterations_log += str("\n\nRegained : "+last_eleminated +␣
↪"\n\n")
                    break
            elif elimination_criteria == "bic":
                criteria = model.bic
                new_model = regressor(y,X)
                new_criteria = new_model.bic
                if criteria < new_criteria:
                    print("Regained : ", last_eleminated)
                    iterations_log += "\n"+str(new_model.summary())+"\nAIC: "+␣
↪str(new_model.aic) + "\nBIC: "+ str(new_model.bic)+"\n"
                    iterations_log += str("\n\nRegained : "+last_eleminated +␣
↪"\n\n")
                    break
            elif elimination_criteria == "adjr2" and model_type =="linear":
                criteria = model.rsquared_adj
                new_model = regressor(y,X)
                new_criteria = new_model.rsquared_adj
                if criteria > new_criteria:
                    print("Regained : ", last_eleminated)
                    iterations_log += "\n"+str(new_model.summary())+"\nAIC: "+␣
↪str(new_model.aic) + "\nBIC: "+ str(new_model.bic)+"\n"
                    iterations_log += str("\n\nRegained : "+last_eleminated +␣
↪"\n\n")
                    break
```

```
            elif elimination_criteria == "r2" and model_type =="linear":
                criteria = model.rsquared
                new_model = regressor(y,X)
                new_criteria = new_model.rsquared
                if criteria > new_criteria:
                    print("Regained : ", last_eleminated)
                    iterations_log += "\n"+str(new_model.summary())+"\nAIC: "+␣
↪str(new_model.aic) + "\nBIC: "+ str(new_model.bic)+"\n"
                    iterations_log += str("\n\nRegained : "+last_eleminated +␣
↪"\n\n")
                    break
            else:
                new_model = regressor(y,X)
            model = new_model
            iterations_log += "\n"+str(model.summary())+"\nAIC: "+ str(model.
↪aic) + "\nBIC: "+ str(model.bic)+"\n"
        else:
            model = regressor(y,X)
            iterations_log += "\n"+str(model.summary())+"\nAIC: "+ str(model.
↪aic) + "\nBIC: "+ str(model.bic)+"\n"
        maxPval = max(model.pvalues)
        cols = X.columns.tolist()
        if maxPval > sl:
            for j in cols:
                if (model.pvalues[j] == maxPval):
                    print("Eliminated :" ,j)
                    iterations_log += str("\n\nEliminated : "+j+ "\n\n")

                    del X[j]
                    last_eleminated = j
        else:
            break
    print(str(model.summary())+"\nAIC: "+ str(model.aic) + "\nBIC: "+ str(model.
↪bic))
    print("Final Variables:", cols)
    iterations_log += "\n"+str(model.summary())+"\nAIC: "+ str(model.aic) +␣
↪"\nBIC: "+ str(model.bic)+"\n"
    return cols, iterations_log
```

Here we prepare the x and y datas to be inserted into the forwardSelection of the script above. This script outputs a bunch of stuff, but what we are looking for is the Final variables list. By taking the variables in that list, we eliminate the weak variables.

[15]:
```
x=dataset[predictors]
y=dataset[['Life Expectancy']]

forwardSelection(x, y)
```

```
Character Variables (Dummies Generated, First Dummies Dropped): []
Entered : Adult Mortality       AIC : 19638.97726850929
Entered : Schooling     AIC : 18093.10638995967
Entered : HIV/AIDS      AIC : 17533.284889162875
Entered : Diphtheria    AIC : 17232.266626851648
Entered : BMI   AIC : 17068.7759565733
Entered : Income Composition of Resources       AIC : 16942.31497374198
Entered : Percentage Expenditure        AIC : 16868.959510401688
Entered : Polio         AIC : 16823.779214008362
Entered : Thinness 1-19 Years   AIC : 16795.972781929326
Entered : Hepatitis B   AIC : 16777.68863832849
Entered : Measles       AIC : 16761.273671339568
Entered : Alcohol       AIC : 16748.116983790605
Entered : Total Expenditure     AIC : 16744.7622682427
Entered : GDP   AIC : 16741.27193367825
Break : Significance Level
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life Expectancy   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.808
Method:                 Least Squares   F-statistic:                     882.4
Date:                Wed, 25 Nov 2020   Prob (F-statistic):               0.00
Time:                        23:07:09   Log-Likelihood:                -8355.6
No. Observations:                2938   AIC:                         1.674e+04
Df Residuals:                    2923   BIC:                         1.683e+04
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
===================
                                    coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
-------------------
intercept                          53.3869      0.587     91.019      0.000
52.237      54.537
Adult Mortality                    -0.0206      0.001    -25.503      0.000
-0.022      -0.019
Schooling                           0.7070      0.043     16.568      0.000
0.623       0.791
HIV/AIDS                           -0.4825      0.018    -26.851      0.000
-0.518      -0.447
Diphtheria                          0.0468      0.005      9.971      0.000
0.038       0.056
BMI                                 0.0424      0.005      8.424      0.000
0.033       0.052
Income Composition of Resources     6.5292      0.649     10.060      0.000
5.257       7.802
Percentage Expenditure              0.0002   8.63e-05      1.802      0.072
```

```
-1.37e-05       0.000
Polio                            0.0321      0.005      7.023      0.000
0.023       0.041
Thinness 1-19 Years             -0.0789      0.022     -3.510      0.000
-0.123      -0.035
Hepatitis B                      -0.0172      0.004     -4.770      0.000
-0.024      -0.010
Measles                       -3.101e-05   6.98e-06     -4.445      0.000
-4.47e-05   -1.73e-05
Alcohol                          0.0861      0.024      3.540      0.000
0.038       0.134
Total Expenditure                0.0883      0.035      2.546      0.011
0.020       0.156
GDP                           3.112e-05   1.33e-05      2.338      0.019
5.02e-06    5.72e-05
==============================================================================
Omnibus:                      132.743   Durbin-Watson:                   0.727
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              378.003
Skew:                          -0.179   Prob(JB):                     8.27e-83
Kurtosis:                       4.720   Cond. No.                     1.33e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.33e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
AIC: 16741.27193367825
BIC: 16831.05419902926
Final Variables: ['intercept', 'Adult Mortality', 'Schooling', 'HIV/AIDS',
'Diphtheria', 'BMI', 'Income Composition of Resources', 'Percentage
Expenditure', 'Polio', 'Thinness 1-19 Years', 'Hepatitis B', 'Measles',
'Alcohol', 'Total Expenditure', 'GDP']
```

[15]:
```
(['intercept',
  'Adult Mortality',
  'Schooling',
  'HIV/AIDS',
  'Diphtheria',
  'BMI',
  'Income Composition of Resources',
  'Percentage Expenditure',
  'Polio',
  'Thinness 1-19 Years',
  'Hepatitis B',
  'Measles',
  'Alcohol',
```

```
  'Total Expenditure',
  'GDP'],
 '\nEntered : Adult Mortality\n\n\n                            OLS Regression
Results                            \n=========================================
==================================\nDep. Variable:        Life Expectancy
R-squared:                       0.482\nModel:                            OLS
Adj. R-squared:                  0.482\nMethod:                 Least Squares
F-statistic:                     2737.\nDate:                Wed, 25 Nov 2020
Prob (F-statistic):               0.00\nTime:                        23:07:09
Log-Likelihood:                -9817.5\nNo. Observations:                2938
AIC:                           1.964e+04\nDf Residuals:                    2936
BIC:                           1.965e+04\nDf Model:                         1
\nCovariance Type:            nonrobust
\n=============================================================================
=====\n                     coef    std err          t      P>|t|      [0.025
0.975]\n-----------------------------------------------------------------------
-----------\nintercept          77.9456      0.209    372.785      0.000
77.536      78.356\nAdult Mortality    -0.0531      0.001    -52.314      0.000
-0.055      -0.051\n=============================================================
==================\nOmnibus:                      1006.278   Durbin-Watson:
0.749\nProb(Omnibus):                  0.000   Jarque-Bera (JB):
3737.589\nSkew:                         -1.678   Prob(JB):
0.00\nKurtosis:                       7.390   Cond. No.
341.\n=============================================================================
====\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\nAIC: 19638.97726850929\nBIC:
19650.948237222758\n\n\nEntered : Schooling\n\n\n                              OLS
Regression Results                             \n================================
==============================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.694\nModel:
OLS   Adj. R-squared:                  0.694\nMethod:                 Least
Squares   F-statistic:                     3334.\nDate:                 Wed, 25
Nov 2020   Prob (F-statistic):               0.00\nTime:
23:07:09   Log-Likelihood:                -9043.6\nNo. Observations:
2938   AIC:                           1.809e+04\nDf Residuals:
2935   BIC:                           1.811e+04\nDf Model:
2                                      \nCovariance Type:
nonrobust                                      \n============================
=====================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n-----------------
-----------------------------------------------------------------\nintercept
57.3533      0.484    118.531      0.000      56.405      58.302\nAdult
Mortality    -0.0363      0.001    -41.931      0.000      -0.038
-0.035\nSchooling           1.4865      0.033     45.118      0.000       1.422
1.551\n=============================================================================
=====\nOmnibus:                       618.728   Durbin-Watson:
0.712\nProb(Omnibus):                  0.000   Jarque-Bera (JB):
```

```
2689.520\nSkew:                          -0.958   Prob(JB):
0.00\nKurtosis:                       7.278   Cond. No.                         1.03
e+03\n==============================================================================
====\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.03e+03.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 18093.10638995967\nBIC: 18111.06284302987\n\n\nEntered :
HIV/AIDS\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.748\nModel:                            OLS   Adj. R-squared:
0.747\nMethod:                 Least Squares   F-statistic:
2897.\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8762.6\nNo. Observations:                2938   AIC:
1.753e+04\nDf Residuals:                    2934   BIC:
1.756e+04\nDf Model:                        3
\nCovariance Type:            nonrobust
\n==============================================================================
=====\n                      coef    std err          t      P>|t|      [0.025
0.975]\n------------------------------------------------------------------------------
-----------\nintercept          56.3567      0.442    127.606      0.000
55.491      57.223\nAdult Mortality    -0.0253      0.001    -28.141      0.000
-0.027      -0.024\nSchooling           1.4936      0.030     49.874      0.000
1.435       1.552\nHIV/AIDS           -0.5069      0.020    -24.865      0.000
-0.547      -0.467\n==========================================================
==================\nOmnibus:                       243.508   Durban-Watson:
0.570\nProb(Omnibus):                  0.000   Jarque-Bera (JB):
1042.703\nSkew:                         -0.295   Prob(JB):
3.80e-227\nKurtosis:                      5.858   Cond. No.
1.04e+03\n==============================================================================
========\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.04e+03.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 17533.284889162875\nBIC: 17557.22682658981\n\n\nEntered :
Diphtheria\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.772\nModel:                            OLS   Adj. R-squared:
0.772\nMethod:                 Least Squares   F-statistic:
2487.\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8611.1\nNo. Observations:                2938   AIC:
1.723e+04\nDf Residuals:                    2933   BIC:
1.726e+04\nDf Model:                        4
\nCovariance Type:            nonrobust
\n==============================================================================
```

```
=====\n                      coef    std err          t      P>|t|      [0.025
0.975]\n------------------------------------------------------------------------------
-----------\nintercept          52.3404      0.476    109.949      0.000
51.407      53.274\nAdult Mortality    -0.0239      0.001    -27.794      0.000
-0.026      -0.022\nSchooling           1.3345      0.030     44.765      0.000
1.276       1.393\nHIV/AIDS           -0.4954      0.019    -25.568      0.000
-0.533      -0.457\nDiphtheria          0.0687      0.004     17.851      0.000
0.061       0.076\n==============================================================
================\nOmnibus:                       169.060   Durbin-Watson:
0.673\nProb(Omnibus):                   0.000   Jarque-Bera (JB):
577.392\nSkew:                          -0.201   Prob(JB):
4.18e-126\nKurtosis:                     5.134   Cond. No.
1.23e+03\n==============================================================
========\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.23e+03.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 17232.266626851648\nBIC: 17262.194048635316\n\n\nEntered :
BMI\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.785\nModel:                            OLS   Adj. R-squared:
0.784\nMethod:                 Least Squares   F-statistic:
2138.\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8528.4\nNo. Observations:                2938   AIC:
1.707e+04\nDf Residuals:                  2932   BIC:
1.710e+04\nDf Model:                       5
\nCovariance Type:            nonrobust
\n==============================================================================
=====\n                      coef    std err          t      P>|t|      [0.025
0.975]\n------------------------------------------------------------------------------
-----------\nintercept          51.7940      0.465    111.434      0.000
50.883      52.705\nAdult Mortality    -0.0222      0.001    -26.284      0.000
-0.024      -0.021\nSchooling           1.1809      0.031     37.739      0.000
1.120       1.242\nHIV/AIDS           -0.4808      0.019    -25.470      0.000
-0.518      -0.444\nDiphtheria          0.0645      0.004     17.153      0.000
0.057       0.072\nBMI                  0.0637      0.005     13.034      0.000
0.054       0.073\n==============================================================
================\nOmnibus:                       132.483   Durbin-Watson:
0.691\nProb(Omnibus):                   0.000   Jarque-Bera (JB):
378.148\nSkew:                          -0.177   Prob(JB):
7.69e-83\nKurtosis:                      4.721   Cond. No.
1.24e+03\n==============================================================
========\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.24e+03.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 17068.7759565733\nBIC: 17104.6888627137\n\n\nEntered : Income
```

```
Composition of Resources\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.794\nModel:                            OLS   Adj. R-squared:
0.794\nMethod:                 Least Squares   F-statistic:
1883.\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8464.2\nNo. Observations:                2938   AIC:
1.694e+04\nDf Residuals:                    2931   BIC:
1.698e+04\nDf Model:                           6
\nCovariance Type:            nonrobust
\n==============================================================================
====================\n                                          coef    std err
t      P>|t|      [0.025      0.975]\n------------------------------------------
-----------------------------------------------------\nintercept
51.3297      0.457    112.410      0.000      50.434      52.225\nAdult
Mortality                     -0.0212      0.001    -25.451      0.000
-0.023      -0.020\nSchooling                               0.8496      0.042
20.161      0.000       0.767       0.932\nHIV/AIDS
-0.4731      0.018    -25.599      0.000      -0.509      -0.437\nDiphtheria
0.0612      0.004     16.583      0.000       0.054       0.068\nBMI
0.0581      0.005     12.086      0.000       0.049       0.068\nIncome
Composition of Resources     7.5536      0.660     11.445      0.000       6.260
8.848\n==============================================================================
=====\nOmnibus:                       139.058   Durbin-Watson:
0.693\nProb(Omnibus):                   0.000   Jarque-Bera (JB):
474.970\nSkew:                          -0.073   Prob(JB):
7.27e-104\nKurtosis:                       4.964   Cond. No.
1.81e+03\n==============================================================================
========\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.81e+03.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 16942.31497374198\nBIC: 16984.213364239116\n\n\nEntered :
Percentage Expenditure\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.799\nModel:                            OLS   Adj. R-squared:
0.799\nMethod:                 Least Squares   F-statistic:
1666.\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8426.5\nNo. Observations:                2938   AIC:
1.687e+04\nDf Residuals:                    2930   BIC:
1.692e+04\nDf Model:                           7
\nCovariance Type:            nonrobust
\n==============================================================================
====================\n                                          coef    std err
t      P>|t|      [0.025      0.975]\n------------------------------------------
```

```
----------------------------------------------------------\nintercept
51.9388      0.456    113.836      0.000      51.044      52.833\nAdult
Mortality                    -0.0206      0.001    -25.029      0.000
-0.022      -0.019\nSchooling                              0.7996      0.042
19.036      0.000       0.717       0.882\nHIV/AIDS
-0.4792      0.018    -26.239      0.000      -0.515      -0.443\nDiphtheria
0.0620      0.004     17.016      0.000       0.055       0.069\nBMI
0.0576      0.005     12.127      0.000       0.048       0.067\nIncome
Composition of Resources     6.8885      0.656     10.499      0.000       5.602
8.175\nPercentage Expenditure              0.0004   4.35e-05      8.725
0.000       0.000       0.000\n==================================================
============================\nOmnibus:                        135.198   Durbin-
Watson:                     0.716\nProb(Omnibus):                  0.000   Jarque-
Bera (JB):              440.488\nSkew:                          -0.099
Prob(JB):                    2.24e-96\nKurtosis:                       4.886
Cond. No.                    1.78e+04\n==========================================
====================================\n\nNotes:\n[1] Standard Errors assume
that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 1.78e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16868.959510401688\nBIC:
16916.843385255557\n\n\nEntered : Polio\n\n\n                            OLS
Regression Results                           \n================================
============================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.802\nModel:
OLS   Adj. R-squared:                  0.802\nMethod:                 Least
Squares   F-statistic:                     1487.\nDate:                Wed, 25
Nov 2020   Prob (F-statistic):              0.00\nTime:
23:07:09   Log-Likelihood:                -8402.9\nNo. Observations:
2938   AIC:                          1.682e+04\nDf Residuals:
2929   BIC:                          1.688e+04\nDf Model:
8                                      \nCovariance Type:
nonrobust                                         \n============================
=====================================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n-----------------
---------------------------------------------------------------------------
--\nintercept                           51.1967      0.465    110.019      0.000
50.284      52.109\nAdult Mortality                      -0.0204      0.001
-24.905      0.000      -0.022      -0.019\nSchooling
0.7779      0.042     18.614      0.000       0.696       0.860\nHIV/AIDS
-0.4786      0.018    -26.414      0.000      -0.514      -0.443\nDiphtheria
0.0428      0.005      9.375      0.000       0.034       0.052\nBMI
0.0560      0.005     11.888      0.000       0.047       0.065\nIncome
Composition of Resources     6.8558      0.651     10.531      0.000       5.579
8.132\nPercentage Expenditure              0.0004   4.31e-05      8.806
0.000       0.000       0.000\nPolio                                  0.0317
0.005      6.886      0.000       0.023       0.041\n==========================
==================================================\nOmnibus:
```

```
126.884   Durbin-Watson:                   0.718\nProb(Omnibus):
0.000   Jarque-Bera (JB):              395.392\nSkew:
-0.095   Prob(JB):                      1.39e-86\nKurtosis:
4.787   Cond. No.                      1.78e+04\n==============================
==============================================\n\nNotes:\n[1] Standard Errors
assume that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 1.78e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16823.779214008362\nBIC:
16877.648573218965\n\n\nEntered : Thinness 1-19 Years\n\n\n
OLS Regression Results                            \n============================
================================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.804\nModel:
OLS   Adj. R-squared:                  0.804\nMethod:                 Least
Squares   F-statistic:                     1338.\nDate:                 Wed, 25
Nov 2020   Prob (F-statistic):              0.00\nTime:
23:07:09   Log-Likelihood:                -8388.0\nNo. Observations:
2938   AIC:                           1.680e+04\nDf Residuals:
2928   BIC:                           1.686e+04\nDf Model:
9                                         \nCovariance Type:
nonrobust                                         \n============================
==============================================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n-----------------
------------------------------------------------------------------------------
--\nintercept                          52.5628      0.526     99.882      0.000
51.531      53.595\nAdult Mortality                     -0.0203      0.001
-24.926      0.000      -0.022      -0.019\nSchooling
0.7507      0.042     17.922      0.000       0.669       0.833\nHIV/AIDS
-0.4741      0.018    -26.267      0.000      -0.510      -0.439\nDiphtheria
0.0423      0.005      9.317      0.000       0.033       0.051\nBMI
0.0458      0.005      9.077      0.000       0.036       0.056\nIncome
Composition of Resources     6.7827      0.648     10.468      0.000       5.512
8.053\nPercentage Expenditure              0.0004   4.31e-05      8.378
0.000       0.000       0.000\nPolio                                 0.0317
0.005      6.924      0.000       0.023       0.041\nThinness 1-19 Years
-0.1184      0.022     -5.464      0.000      -0.161      -0.076\n==============
==============================================================\nOmnibus:
127.160   Durbin-Watson:                   0.707\nProb(Omnibus):
0.000   Jarque-Bera (JB):              366.203\nSkew:
-0.154   Prob(JB):                      3.02e-80\nKurtosis:
4.702   Cond. No.                      1.79e+04\n==============================
==============================================\n\nNotes:\n[1] Standard Errors
assume that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 1.79e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16795.972781929326\nBIC:
16855.827625496662\n\n\nEntered : Hepatitis B\n\n\n
OLS Regression Results                            \n============================
================================================\nDep. Variable:        Life
```

```
Expectancy   R-squared:                       0.806\nModel:
OLS   Adj. R-squared:                  0.805\nMethod:                 Least
Squares   F-statistic:                     1214.\nDate:                 Wed, 25
Nov 2020   Prob (F-statistic):              0.00\nTime:
23:07:09   Log-Likelihood:                -8377.8\nNo. Observations:
2938   AIC:                           1.678e+04\nDf Residuals:
2927   BIC:                           1.684e+04\nDf Model:
10                                         \nCovariance Type:
nonrobust                                         \n==============================
==============================================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n------------------
------------------------------------------------------------------------------
--\nintercept                            53.4309      0.559     95.612      0.000
52.335      54.527\nAdult Mortality                     -0.0203      0.001
-25.060      0.000      -0.022      -0.019\nSchooling
0.7499      0.042     17.963      0.000       0.668       0.832\nHIV/AIDS
-0.4749      0.018    -26.393      0.000      -0.510      -0.440\nDiphtheria
0.0482      0.005     10.227      0.000       0.039       0.057\nBMI
0.0456      0.005      9.051      0.000       0.036       0.055\nIncome
Composition of Resources     6.6256      0.647     10.244      0.000       5.357
7.894\nPercentage Expenditure              0.0004   4.29e-05      8.479
0.000       0.000       0.000\nPolio                                  0.0336
0.005      7.328      0.000       0.025       0.043\nThinness 1-19 Years
-0.1200      0.022     -5.557      0.000      -0.162      -0.078\nHepatitis B
-0.0163      0.004     -4.503      0.000      -0.023      -0.009\n==============
==============================================================\nOmnibus:
121.453   Durbin-Watson:                   0.720\nProb(Omnibus):
0.000   Jarque-Bera (JB):              353.409\nSkew:
-0.124   Prob(JB):                     1.81e-77\nKurtosis:
4.681   Cond. No.                      1.80e+04\n================================
============================================\n\nNotes:\n[1] Standard Errors
assume that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 1.8e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16777.68863832849\nBIC:
16843.528966252565\n\n\nEntered : Measles\n\n\n                                OLS
Regression Results                              \n================================
============================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.807\nModel:
OLS   Adj. R-squared:                  0.806\nMethod:                 Least
Squares   F-statistic:                     1112.\nDate:                 Wed, 25
Nov 2020   Prob (F-statistic):              0.00\nTime:
23:07:09   Log-Likelihood:                -8368.6\nNo. Observations:
2938   AIC:                           1.676e+04\nDf Residuals:
2926   BIC:                           1.683e+04\nDf Model:
11                                         \nCovariance Type:
nonrobust                                         \n==============================
==============================================================================\n
```

```
coef    std err          t      P>|t|      [0.025      0.975]\n------------------
------------------------------------------------------------------------------
--\nintercept                           53.6616      0.560     95.864      0.000
52.564      54.759\nAdult Mortality                    -0.0206      0.001
-25.363      0.000      -0.022      -0.019\nSchooling
0.7518      0.042     18.060      0.000       0.670       0.833\nHIV/AIDS
-0.4749      0.018    -26.474      0.000      -0.510      -0.440\nDiphtheria
0.0472      0.005     10.041      0.000       0.038       0.056\nBMI
0.0442      0.005      8.784      0.000       0.034       0.054\nIncome
Composition of Resources     6.6058      0.645     10.243      0.000       5.341
7.870\nPercentage Expenditure              0.0004   4.28e-05      8.516
0.000       0.000       0.000\nPolio                                0.0328
0.005      7.172      0.000       0.024       0.042\nThinness 1-19 Years
-0.1056      0.022     -4.847      0.000      -0.148      -0.063\nHepatitis B
-0.0163      0.004     -4.524      0.000      -0.023      -0.009\nMeasles
-2.993e-05   6.98e-06     -4.289      0.000   -4.36e-05   -1.62e-05\n==========
==============================================================\nOmnibus:
121.725   Durbin-Watson:                   0.727\nProb(Omnibus):
0.000   Jarque-Bera (JB):              356.229\nSkew:
-0.121   Prob(JB):                     4.42e-78\nKurtosis:
4.689   Cond. No.                     9.95e+04\n================================
==============================================\n\nNotes:\n[1] Standard Errors
assume that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 9.95e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16761.273671339568\nBIC:
16833.099483620375\n\n\nEntered : Alcohol\n\n\n                                OLS
Regression Results                            \n================================
==============================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.808\nModel:
OLS   Adj. R-squared:                  0.807\nMethod:                 Least
Squares   F-statistic:                     1026.\nDate:                 Wed, 25
Nov 2020   Prob (F-statistic):              0.00\nTime:
23:07:09   Log-Likelihood:                -8361.1\nNo. Observations:
2938   AIC:                          1.675e+04\nDf Residuals:
2925   BIC:                          1.683e+04\nDf Model:
12                                       \nCovariance Type:
nonrobust                                     \n============================
==============================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n------------------
------------------------------------------------------------------------------
--\nintercept                           53.7475      0.559     96.174      0.000
52.652      54.843\nAdult Mortality                    -0.0207      0.001
-25.587      0.000      -0.022      -0.019\nSchooling
0.7134      0.043     16.712      0.000       0.630       0.797\nHIV/AIDS
-0.4803      0.018    -26.759      0.000      -0.516      -0.445\nDiphtheria
0.0471      0.005     10.037      0.000       0.038       0.056\nBMI
0.0441      0.005      8.784      0.000       0.034       0.054\nIncome
```

```
Composition of Resources     6.5769      0.643     10.223      0.000       5.315
7.838\nPercentage Expenditure              0.0003   4.33e-05      7.761
0.000       0.000       0.000\nPolio                                  0.0325
0.005      7.126      0.000       0.024       0.042\nThinness 1-19 Years
-0.0846      0.022     -3.780      0.000      -0.129      -0.041\nHepatitis B
-0.0168      0.004     -4.662      0.000      -0.024      -0.010\nMeasles
-3.187e-05   6.98e-06     -4.566      0.000   -4.56e-05   -1.82e-05\nAlcohol
0.0934      0.024      3.890      0.000       0.046       0.140\n==============
================================================================\nOmnibus:
123.168   Durbin-Watson:                   0.727\nProb(Omnibus):
0.000   Jarque-Bera (JB):              343.388\nSkew:
-0.158   Prob(JB):                     2.72e-75\nKurtosis:
4.645   Cond. No.                     9.96e+04\n================================
================================================\n\nNotes:\n[1] Standard Errors
assume that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 9.96e+04. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16748.116983790605\nBIC:
16825.928280428147\n\n\nEntered : Total Expenditure\n\n\n
OLS Regression Results                                \n============================
================================================\nDep. Variable:        Life
Expectancy   R-squared:                       0.808\nModel:
OLS   Adj. R-squared:                  0.807\nMethod:                 Least
Squares   F-statistic:                      948.5\nDate:                Wed, 25
Nov 2020   Prob (F-statistic):               0.00\nTime:
23:07:09   Log-Likelihood:                -8358.4\nNo. Observations:
2938   AIC:                          1.674e+04\nDf Residuals:
2924   BIC:                          1.683e+04\nDf Model:
13                                       \nCovariance Type:
nonrobust                                       \n============================
=====================================================================\n
coef    std err          t      P>|t|      [0.025      0.975]\n-----------------
---------------------------------------------------------------------------
--\nintercept                            53.3332      0.587     90.928      0.000
52.183      54.483\nAdult Mortality                    -0.0207      0.001
-25.517      0.000      -0.022      -0.019\nSchooling
0.7091      0.043     16.609      0.000       0.625       0.793\nHIV/AIDS
-0.4832      0.018    -26.875      0.000      -0.518      -0.448\nDiphtheria
0.0466      0.005      9.936      0.000       0.037       0.056\nBMI
0.0431      0.005      8.577      0.000       0.033       0.053\nIncome
Composition of Resources     6.7033      0.645     10.389      0.000       5.438
7.968\nPercentage Expenditure              0.0003   4.34e-05      7.607
0.000       0.000       0.000\nPolio                                  0.0325
0.005      7.123      0.000       0.024       0.041\nThinness 1-19 Years
-0.0793      0.022     -3.524      0.000      -0.123      -0.035\nHepatitis B
-0.0167      0.004     -4.650      0.000      -0.024      -0.010\nMeasles
-3.111e-05   6.98e-06     -4.455      0.000   -4.48e-05   -1.74e-05\nAlcohol
0.0840      0.024      3.452      0.001       0.036       0.132\nTotal
```

```
Expenditure                    0.0797      0.035      2.310      0.021
0.012       0.147\n==============================================================
================\nOmnibus:                       131.446   Durbin-Watson:
0.728\nProb(Omnibus):                  0.000   Jarque-Bera (JB):
378.391\nSkew:                          -0.169   Prob(JB):
6.82e-83\nKurtosis:                        4.725   Cond. No.
1.01e+05\n==============================================================================
========\n\nNotes:\n[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.\n[2] The condition number is large, 1.01e+05.
This might indicate that there are\nstrong multicollinearity or other numerical
problems.\nAIC: 16744.7622682427\nBIC: 16828.559049236974\n\n\nEntered :
GDP\n\n\n                            OLS Regression Results
\n==============================================================================
\nDep. Variable:        Life Expectancy   R-squared:
0.809\nModel:                            OLS   Adj. R-squared:
0.808\nMethod:                 Least Squares   F-statistic:
882.4\nDate:                Wed, 25 Nov 2020   Prob (F-statistic):
0.00\nTime:                        23:07:09   Log-Likelihood:
-8355.6\nNo. Observations:                2938   AIC:
1.674e+04\nDf Residuals:                    2923   BIC:
1.683e+04\nDf Model:                          14
\nCovariance Type:            nonrobust
\n==============================================================================
====================\n                                        coef    std err
t      P>|t|      [0.025      0.975]\n------------------------------------------
-------------------------------------------------------\nintercept
53.3869      0.587     91.019      0.000      52.237      54.537\nAdult
Mortality                    -0.0206      0.001    -25.503      0.000
-0.022      -0.019\nSchooling                             0.7070      0.043
16.568      0.000       0.623       0.791\nHIV/AIDS
-0.4825      0.018    -26.851      0.000      -0.518      -0.447\nDiphtheria
0.0468      0.005      9.971      0.000       0.038       0.056\nBMI
0.0424      0.005      8.424      0.000       0.033       0.052\nIncome
Composition of Resources     6.5292      0.649     10.060      0.000       5.257
7.802\nPercentage Expenditure              0.0002   8.63e-05      1.802
0.072   -1.37e-05       0.000\nPolio                                0.0321
0.005      7.023      0.000       0.023       0.041\nThinness 1-19 Years
-0.0789      0.022     -3.510      0.000      -0.123      -0.035\nHepatitis B
-0.0172      0.004     -4.770      0.000      -0.024      -0.010\nMeasles
-3.101e-05   6.98e-06     -4.445      0.000   -4.47e-05   -1.73e-05\nAlcohol
0.0861      0.024      3.540      0.000       0.038       0.134\nTotal
Expenditure                    0.0883      0.035      2.546      0.011
0.020       0.156\nGDP                                3.112e-05   1.33e-05
2.338      0.019    5.02e-06    5.72e-05\n=====================================
=========================================\nOmnibus:                       132.743
Durbin-Watson:                   0.727\nProb(Omnibus):                  0.000
Jarque-Bera (JB):              378.003\nSkew:                          -0.179
```

```
Prob(JB):                     8.27e-83\nKurtosis:                       4.720
Cond. No.                     1.33e+05\n==========================================
==========================================\n\nNotes:\n[1] Standard Errors assume
that the covariance matrix of the errors is correctly specified.\n[2] The
condition number is large, 1.33e+05. This might indicate that there are\nstrong
multicollinearity or other numerical problems.\nAIC: 16741.27193367825\nBIC:
16831.05419902926\n\n')
```

In here, we run the Multiple Linear Regression with the deleted weak variables and get the new adjusted rsquared.

[16]:
```
x = dataset[['Adult Mortality', 'Schooling', 'HIV/AIDS', 'Diphtheria', 'BMI',␣
 ↪'Income Composition of Resources', 'Percentage Expenditure', 'Polio',␣
 ↪'Thinness 1-19 Years', 'Hepatitis B', 'Measles', 'Alcohol', 'Total␣
 ↪Expenditure', 'GDP']]
y = dataset[['Life Expectancy']]

X = sm.add_constant(x)
optimized_reg = sm.OLS(y, X).fit()

optimized_reg.rsquared_adj
```

[16]:
```
0.8077530219116467
```

[ ]:
```

```