Life Expectancy Prediction

MACHINE LEARNING MODEL

- Satyam Mishra

Life Expectancy

Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age, and other demographic factors including sex. The most commonly used measure is **life expectancy at birth** (**LEB**), which can be defined in two ways. *Cohort* LEB is the mean length of life of an actual birth cohort (all individuals born in a given year) and can be computed only for cohorts born many decades ago so that all their members have died. *Period* LEB is the mean length of life of a hypothetical cohort assumed to be exposed, from birth through death, to the mortality rates observed at a given year.
National LEB figures reported by national agencies and international organizations for human populations are indeed estimates of *period* LEB. In the Bronze Age and the Iron Age, human LEB was 26 years; the 2010 world LEB was 67.2 years. In recent years, LEB in Eswatini (Swaziland) is about 49, while LEB in Japan is about 83. The combination of high infant mortality and deaths in young adulthood from accidents, epidemics, plagues, wars, and childbirth, particularly before modern medicine was widely available, significantly lowers LEB. For example, a society with a LEB of 40 may have few people dying at precisely 40: most will die before 30 or after 55. In populations with high infant mortality rates, LEB is highly sensitive to the rate of death in the first few years of life. Because of this sensitivity to infant mortality, LEB can be subjected to gross misinterpretation, leading one to believe that a population with a low LEB will necessarily have a small proportion of older people. Another measure, such as life expectancy at age 5 (e_5), can be used to exclude the effect of infant mortality to provide a simple measure of overall mortality rates other than in early childhood; in the hypothetical population above, life expectancy at 5 would be another 65.Aggregate population measures, such as the proportion of the population in various age groups, should also be used alongside individual-based measures like formal life expectancy when analyzing population structure and dynamics. However, pre-modern societies still had universally higher mortality rates and universally lower life expectancies at every age for both genders, and this example was relatively rare. In societies with life

expectancies of 30, for instance, a 40-year remaining timespan at age 5 may not have been uncommon, but a 60-year one was.

Factors Affecting Life Expectancy

Economic development and the improvement in some environmental conditions (for example in many urban areas), improved lifestyles, advances in healthcare and medicine, including reduced infant mortality, have resulted in a continuous increase in life expectancy at birth during the last century.
Significant factors in life expectancy include gender, genetics, access to health care, hygiene, diet and nutrition, exercise, lifestyle, and crime rates.
Evidence-based studies indicate that longevity is based on two major factors, genetics and lifestyle choices.

Introduction

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Approach Used

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

Source

The data was collected from WHO and United Nation Website

Website Link :- http://data.un.org/data.aspx?d=popdiv&f=variableID%3A68

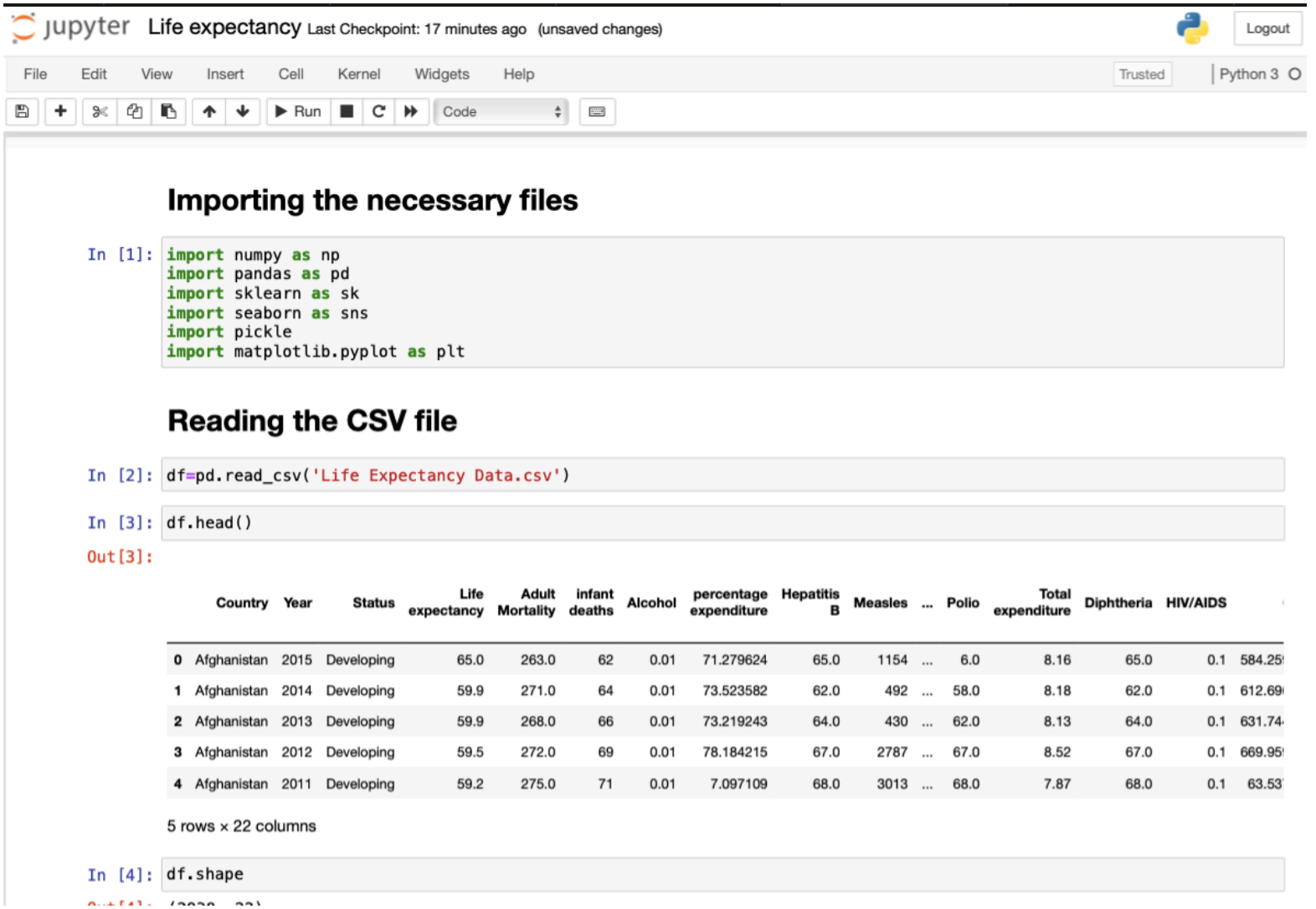
Multi Variable regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

What Multiple Linear Regression Can Tell You??

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

MODEL



Importing the necessary files

In [1]:
```
import numpy as np
import pandas as pd
import sklearn as sk
import seaborn as sns
import pickle
import matplotlib.pyplot as plt
```

Reading the CSV file

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

In [3]:
```
df.head()
```

Out[3]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.25
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.69
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.74
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.95
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.53

5 rows × 22 columns

In [4]:
```
df.shape
```

jupyter Life expectancy Last Checkpoint: 19 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [4]:
```
df.shape
```

Out[4]: (2938, 22)

Checking for null values

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
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
```

Replacing NUll values with their respective mean

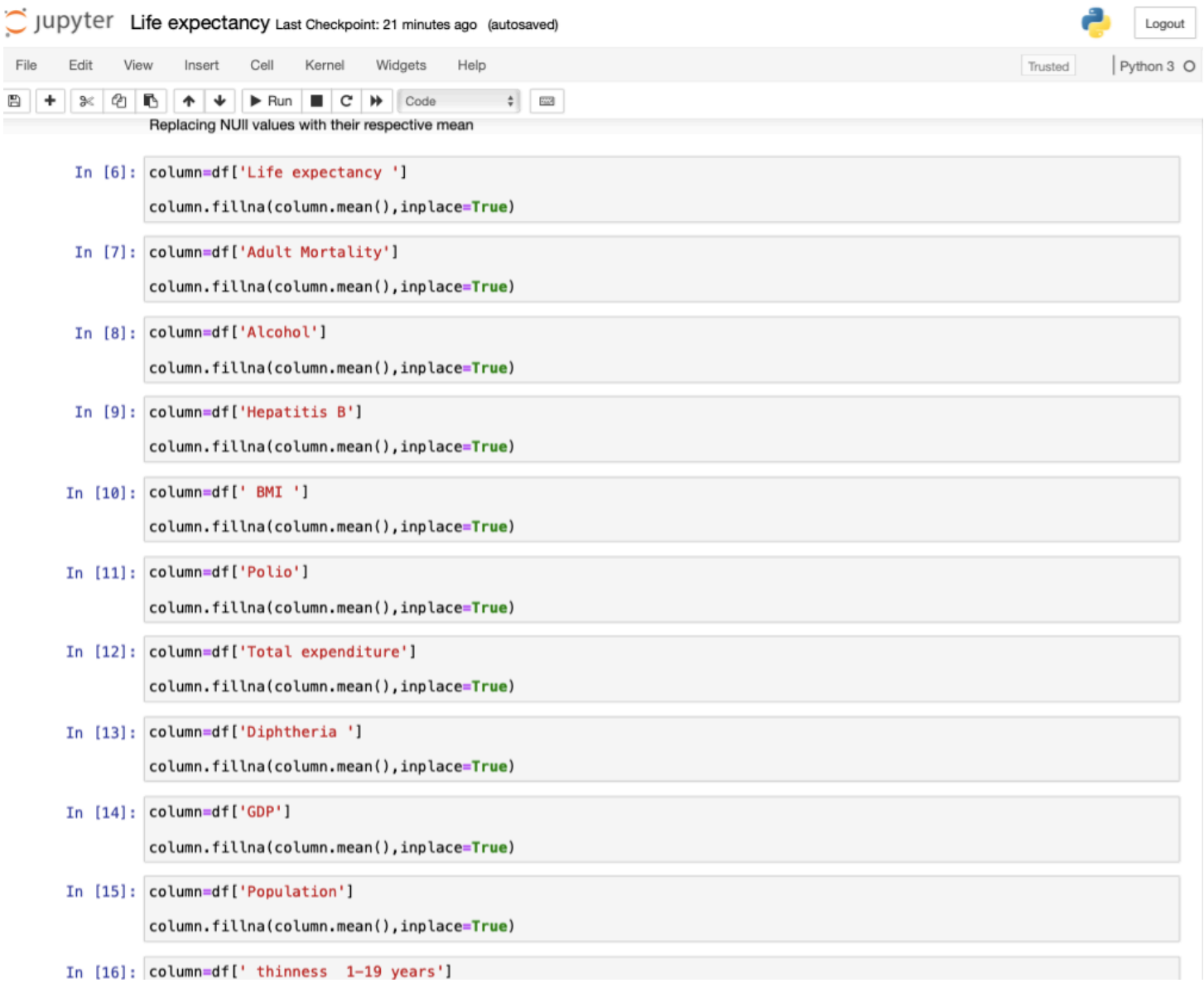
In [6]:
```
column=df['Life expectancy ']

column.fillna(column.mean(),inplace=True)
```

In [7]:
```
column=df['Adult Mortality']

column.fillna(column.mean(),inplace=True)
```

In [8]:
```
column=df['Alcohol']
```


jupyter Life expectancy Last Checkpoint: 21 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Replacing NUll values with their respective mean
In [6]: column=df['Life expectancy ']
column.fillna(column.mean(),inplace=True)
In [7]: column=df['Adult Mortality']
column.fillna(column.mean(),inplace=True)
In [8]: column=df['Alcohol']
column.fillna(column.mean(),inplace=True)
In [9]: column=df['Hepatitis B']
column.fillna(column.mean(),inplace=True)
In [10]: column=df[' BMI ']
column.fillna(column.mean(),inplace=True)
In [11]: column=df['Polio']
column.fillna(column.mean(),inplace=True)
In [12]: column=df['Total expenditure']
column.fillna(column.mean(),inplace=True)
In [13]: column=df['Diphtheria ']
column.fillna(column.mean(),inplace=True)
In [14]: column=df['GDP']
column.fillna(column.mean(),inplace=True)
In [15]: column=df['Population']
column.fillna(column.mean(),inplace=True)
In [16]: column=df[' thinness 1-19 years']

jupyter Life expectancy Last Checkpoint: 22 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Code

In [16]:
```
column=df[' thinness  1-19 years']

column.fillna(column.mean(),inplace=True)
```

In [17]:
```
column=df[' thinness 5-9 years']

column.fillna(column.mean(),inplace=True)
```

In [18]:
```
column=df['Income composition of resources']

column.fillna(column.mean(),inplace=True)
```

In [19]:
```
column=df['Schooling']

column.fillna(column.mean(),inplace=True)
```

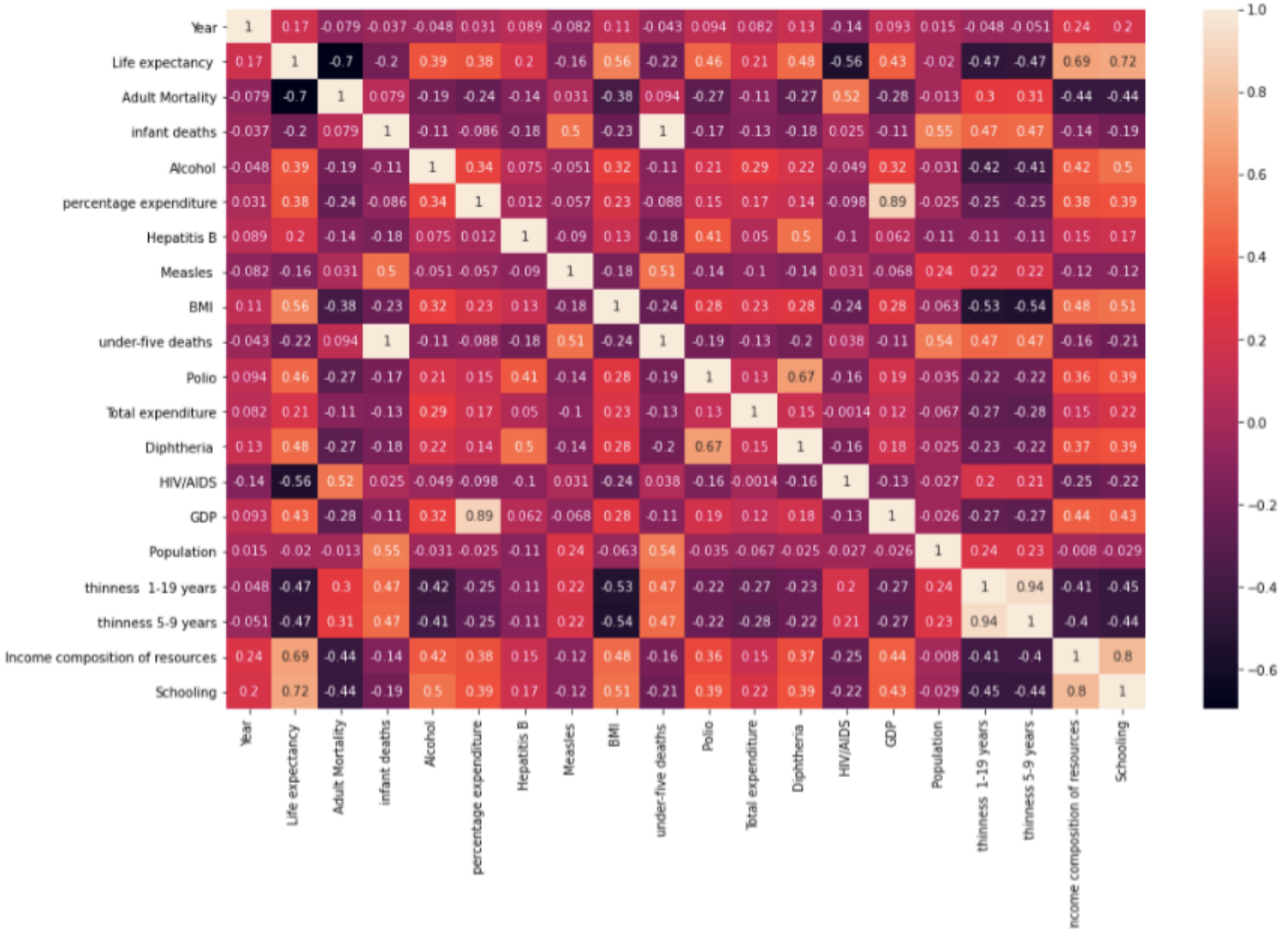
In [20]:
```
df.isnull().sum()
```

Out[20]:
```
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
```

NULL values are removed

```
26]: # Correlation heatmap
     plt.figure(figsize=(16,10))
     sns.heatmap(df.corr(),annot=True)
```

```
26]: <AxesSubplot:>
```



upyter Life expectancy Last Checkpoint: 25 minutes ago (autosaved)

Logout

Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [25]:
```
import statsmodels.api as sm
X_train_sm = x
# statsmodel dont include β0 so you need to add sm.add_constant(X) in order to add a constant
X_train_sm = sm.add_constant(X_train_sm)
# now creating a fitted model in one line
lm_1 = sm.OLS(y,X_train_sm).fit()

# print the coefficients
print(lm_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Life expectancy   R-squared:                       0.804
Model:                          OLS   Adj. R-squared:                  0.802
Method:               Least Squares   F-statistic:                     746.8
Date:              Wed, 29 Sep 2021   Prob (F-statistic):               0.00
Time:                      22:04:38   Log-Likelihood:                -8394.4
No. Observations:              2938   AIC:                         1.682e+04
Df Residuals:                  2921   BIC:                         1.692e+04
Df Model:                        16
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                             54.3852      0.574     94.728      0.000      53.259      55.511
Adult Mortality                   -0.0216      0.001    -26.362      0.000      -0.023      -0.020
infant deaths                     -0.0010      0.001     -0.989      0.323      -0.003       0.001
Alcohol                            0.0784      0.025      3.171      0.002       0.030       0.127
percentage expenditure          9.323e-05   8.75e-05      1.066      0.287   -7.83e-05       0.000
Hepatitis B                       -0.0167      0.004     -4.088      0.000      -0.025      -0.009
Measles                        -3.395e-05   7.92e-06     -4.284      0.000   -4.95e-05   -1.84e-05
Polio                              0.0334      0.005      7.188      0.000       0.024       0.043
Total expenditure                  0.1196      0.035      3.407      0.001       0.051       0.188
Diphtheria                         0.0474      0.005      9.737      0.000       0.038       0.057
 HIV/AIDS                         -0.4884      0.018    -26.739      0.000      -0.524      -0.453
GDP                             3.858e-05   1.35e-05      2.863      0.004    1.22e-05     6.5e-05
Population                      3.01e-09   1.75e-09      1.719      0.086   -4.23e-10    6.44e-09
 thinness  1-19 years            -0.1241      0.052     -2.366      0.018      -0.227      -0.021
 thinness 5-9 years              -0.0224      0.051     -0.436      0.663      -0.123       0.078
Income composition of resources    7.1192      0.656     10.846      0.000       5.832       8.406
Schooling                          0.7354      0.043     17.065      0.000       0.651       0.820
==============================================================================
Omnibus:                      163.839   Durbin-Watson:                   0.711
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              509.771
Skew:                          -0.231   Prob(JB):                    2.02e-111
Kurtosis:                       4.988   Cond. No.                     4.78e+08
==============================================================================
```

```
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     746.8
Date:                Wed, 29 Sep 2021   Prob (F-statistic):               0.00
Time:                        22:04:38   Log-Likelihood:                -8394.4
No. Observations:                2938   AIC:                         1.682e+04
Df Residuals:                    2921   BIC:                         1.692e+04
Df Model:                          16
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           54.3852      0.574     94.728      0.000      53.259      55.511
Adult Mortality                 -0.0216      0.001    -26.362      0.000      -0.023      -0.020
infant deaths                   -0.0010      0.001     -0.989      0.323      -0.003       0.001
Alcohol                          0.0784      0.025      3.171      0.002       0.030       0.127
percentage expenditure        9.323e-05   8.75e-05      1.066      0.287   -7.83e-05       0.000
Hepatitis B                     -0.0167      0.004     -4.088      0.000      -0.025      -0.009
Measles                      -3.395e-05   7.92e-06     -4.284      0.000   -4.95e-05   -1.84e-05
Polio                            0.0334      0.005      7.188      0.000       0.024       0.043
Total expenditure                0.1196      0.035      3.407      0.001       0.051       0.188
Diphtheria                       0.0474      0.005      9.737      0.000       0.038       0.057
 HIV/AIDS                       -0.4884      0.018    -26.739      0.000      -0.524      -0.453
GDP                           3.858e-05   1.35e-05      2.863      0.004    1.22e-05     6.5e-05
Population                    3.01e-09   1.75e-09      1.719      0.086   -4.23e-10    6.44e-09
 thinness  1-19 years           -0.1241      0.052     -2.366      0.018      -0.227      -0.021
 thinness 5-9 years             -0.0224      0.051     -0.436      0.663      -0.123       0.078
Income composition of resources   7.1192      0.656     10.846      0.000       5.832       8.406
Schooling                         0.7354      0.043     17.065      0.000       0.651       0.820
==============================================================================
Omnibus:                      163.839   Durbin-Watson:                   0.711
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              509.771
Skew:                          -0.231   Prob(JB):                    2.02e-111
Kurtosis:                       4.988   Cond. No.                     4.78e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.78e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The above Stats show that even our model is giving R^2 value 80.4% but there are strong multicolinear problems .
'thinness 5-9 years', 'thinness 1-19 years ','percentage expenditure' ,'infant deaths' has p values greater than > 0.05
All of the above this that there are strong **Multicolinear problem** . So in order to remove that we will use a technique called **Recursive feature Elimination (RFE) It is an efficient approach for eliminating features from a training dataset for feature selection.**

In [ ]:

jupyter Life expectancy Last Checkpoint: 29 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

In [ ]:

RFE technique

In [27]:
```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
```

In [28]:
```
lr=LinearRegression()
```

In [29]:
```
rfe=RFE(lr,10)
rfe=rfe.fit(x,y)
print(rfe.support_)
print(rfe.ranking_)
```

```
[ True False  True False False False  True  True  True  True False False
  True  True  True  True]
[1 3 1 4 2 6 1 1 1 1 5 7 1 1 1 1]
```

```
/Users/satyammishra/opt/anaconda3/lib/python3.8/site-packages/sklearn/utils/validation.py:70: FutureWarnin
n_features_to_select=10 as keyword args. From version 1.0 (renaming of 0.25) passing these as positional a
will result in an error
  warnings.warn(f"Pass {args_msg} as keyword args. From version "
```

The above outcome is suggesting which columns should be selected to keep the **multicolinear problem as minimum as possible**

Creating a regression model by only considering the required values

In [36]:
```
X=df[['Adult Mortality','Alcohol','Polio','Total expenditure','Diphtheria ',' HIV/AIDS','Income compositio
Y=df['Life expectancy ']
lr.fit(X,Y)
```

Out[36]:
```
LinearRegression()
```

In [37]:
```
lr.intercept_
```

Out[37]:
```
50.98784152307667
```

In [38]:
```
lr.coef_
```

Out[38]:
```
array([-0.02228954,  0.13992066,  0.03293915,  0.18021484,  0.04173663,
       -0.49836421,  8.38135222,  0.80461282])
```

pyter **Life expectancy** Last Checkpoint: 31 minutes ago (autosaved)

Logout

Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [37]:
```
lr.intercept_
```

Out[37]:
```
50.98784152307667
```

In [38]:
```
lr.coef_
```

Out[38]:
```
array([-0.02228954,  0.13992066,  0.03293915,  0.18021484,  0.04173663,
       -0.49836421,  8.38135222,  0.80461282])
```

Coefficients of every features 👆

In [39]:
```
import statsmodels.api as sm
x_train_sm=X
x_train_sm=sm.add_constant(x_train_sm)
lm1=sm.OLS(Y,x_train_sm).fit()
print(lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Life expectancy   R-squared:                       0.792
Model:                          OLS   Adj. R-squared:                  0.791
Method:               Least Squares   F-statistic:                     1394.
Date:              Wed, 29 Sep 2021   Prob (F-statistic):               0.00
Time:                      22:55:17   Log-Likelihood:                -8478.6
No. Observations:              2938   AIC:                         1.698e+04
Df Residuals:                  2929   BIC:                         1.703e+04
Df Model:                         8
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                             50.9878      0.501    101.857      0.000      50.006      51.969
Adult Mortality                   -0.0223      0.001    -26.747      0.000      -0.024      -0.021
Alcohol                            0.1399      0.024      5.764      0.000       0.092       0.188
Polio                              0.0329      0.005      6.962      0.000       0.024       0.042
Total expenditure                  0.1802      0.035      5.111      0.000       0.111       0.249
Diphtheria                         0.0417      0.005      8.888      0.000       0.033       0.051
 HIV/AIDS                         -0.4984      0.019    -26.765      0.000      -0.535      -0.462
Income composition of resources    8.3814      0.662     12.667      0.000       7.084       9.679
Schooling                          0.8046      0.044     18.364      0.000       0.719       0.891
==============================================================================
Omnibus:                      178.367   Durbin-Watson:                   0.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              663.009
Skew:                          -0.184   Prob(JB):                    1.07e-144
Kurtosis:                       5.298   Cond. No.                     1.89e+03
==============================================================================
```

Now 'P' value of every column is less than 0.05 which seems good

In [40]:
```
pickle.dump(lr,open('model.pkl','wb'))

model=pickle.load(open('model.pkl','rb'))
```

In [43]:
```
x=model.predict([[56.0,67,58,8.16,8,0.1,.479,1]])
```

In [45]:
```
x
```

Out[45]:
```
array([67.5986724])
```

Conclusions Drawn

With the help of RFE technique the features which shows least intercorrelation are Adult Mortality, Alcohol, Polio, Total expenditure , Diphtheria, HIV / AIDS, Income composition of resources, Schooling (tot years)