---Team D - Make Logic Mahaveeer Chand Rushvi Jain Srikanth Lakkoju

Data set - Life Expectancy from Year 2000 - 2015

In [50]:
```
# Import basic libraries
import numpy as np
import pandas as pd
# import visualization libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [51]:
```
# Load the CSV file - dataset
df = pd.read_csv('Life Expectancy Data.csv')
df.head(5)
```

Out[51]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	M
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	

5 rows × 22 columns

In [52]:
```
# The dataset has 2938 rows and 22 columns
print("Shape of the dataset ", df.shape)
```

```
Shape of the dataset  (2938, 22)
```

In [53]:
```
# Below command shows the non-null count of the variables and the datatypes.
# Verifying whether data for each variable is according to its dataype or not.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Year                             2938 non-null   int64
 2   Status                           2938 non-null   object
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64
```

```
 10   BMI                              2904 non-null   float64
 11  under-five deaths                 2938 non-null   int64
 12  Polio                             2919 non-null   float64
 13  Total expenditure                 2712 non-null   float64
 14  Diphtheria                        2919 non-null   float64
 15   HIV/AIDS                         2938 non-null   float64
 16  GDP                               2490 non-null   float64
 17  Population                        2286 non-null   float64
 18   thinness  1-19 years             2904 non-null   float64
 19   thinness 5-9 years               2904 non-null   float64
 20  Income composition of resources   2771 non-null   float64
 21  Schooling                         2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

In [54]:
```
#Statistical analysis of the features
df.describe().T
```

Out[54]:

	count	mean	std	min	25%	50%	
Year	2938.0	2.007519e+03	4.613841e+00	2000.00000	2004.000000	2.008000e+03	2.01200(
Life expectancy	2928.0	6.922493e+01	9.523867e+00	36.30000	63.100000	7.210000e+01	7.57000(
Adult Mortality	2928.0	1.647964e+02	1.242921e+02	1.00000	74.000000	1.440000e+02	2.28000(
infant deaths	2938.0	3.030395e+01	1.179265e+02	0.00000	0.000000	3.000000e+00	2.20000(
Alcohol	2744.0	4.602861e+00	4.052413e+00	0.01000	0.877500	3.755000e+00	7.70250(
percentage expenditure	2938.0	7.382513e+02	1.987915e+03	0.00000	4.685343	6.491291e+01	4.41534(
Hepatitis B	2385.0	8.094046e+01	2.507002e+01	1.00000	77.000000	9.200000e+01	9.70000(
Measles	2938.0	2.419592e+03	1.146727e+04	0.00000	0.000000	1.700000e+01	3.60250(
BMI	2904.0	3.832125e+01	2.004403e+01	1.00000	19.300000	4.350000e+01	5.62000(
under-five deaths	2938.0	4.203574e+01	1.604455e+02	0.00000	0.000000	4.000000e+00	2.80000(
Polio	2919.0	8.255019e+01	2.342805e+01	3.00000	78.000000	9.300000e+01	9.70000(
Total expenditure	2712.0	5.938190e+00	2.498320e+00	0.37000	4.260000	5.755000e+00	7.49250(
Diphtheria	2919.0	8.232408e+01	2.371691e+01	2.00000	78.000000	9.300000e+01	9.70000(
HIV/AIDS	2938.0	1.742103e+00	5.077785e+00	0.10000	0.100000	1.000000e-01	8.00000
GDP	2490.0	7.483158e+03	1.427017e+04	1.68135	463.935626	1.766948e+03	5.91080(
Population	2286.0	1.275338e+07	6.101210e+07	34.00000	195793.250000	1.386542e+06	7.42035(
thinness 1-19 years	2904.0	4.839704e+00	4.420195e+00	0.10000	1.600000	3.300000e+00	7.20000(
thinness 5-9 years	2904.0	4.870317e+00	4.508882e+00	0.10000	1.500000	3.300000e+00	7.20000(
Income composition of resources	2771.0	6.275511e-01	2.109036e-01	0.00000	0.493000	6.770000e-01	7.79000
Schooling	2775.0	1.199279e+01	3.358920e+00	0.00000	10.100000	1.230000e+01	1.43000(

In [55]:
```
df.rename(columns={" BMI ":"BMI","Life expectancy ":"Life_Expectancy","Adult Mortali
                   "infant deaths":"Infant_Deaths","percentage expenditure":"Percent
                  "Measles ":"Measles"," BMI ":"BMI","under-five deaths ":"Under_Fiv
                  " HIV/AIDS":"HIV/AIDS"," thinness  1-19 years":"thinness_1to19_yea
                   "Total expenditure":"Tot_Exp"},inplace=True)
```

In [56]:
```
# Identify percentage of null values in each column.
df.isnull().sum()*100/df.isnull().count()
```

Out[56]:
```
Country                          0.000000
Year                             0.000000
Status                           0.000000
Life_Expectancy                  0.340368
Adult_Mortality                  0.340368
Infant_Deaths                    0.000000
Alcohol                          6.603131
Percentage_Exp                   0.000000
HepatitisB                      18.822328
Measles                          0.000000
BMI                              1.157250
Under_Five_Deaths                0.000000
Polio                            0.646698
Tot_Exp                          7.692308
Diphtheria                       0.646698
HIV/AIDS                         0.000000
GDP                             15.248468
Population                      22.191967
thinness_1to19_years             1.157250
thinness_5to9_years              1.157250
Income_Comp_Of_Resources         5.684139
Schooling                        5.547992
dtype: float64
```

In [57]:
```
# After renaming the columns by removing trailing spaces. Identify percentage of nul
df.isnull().sum()*100/df.isnull().count()
```

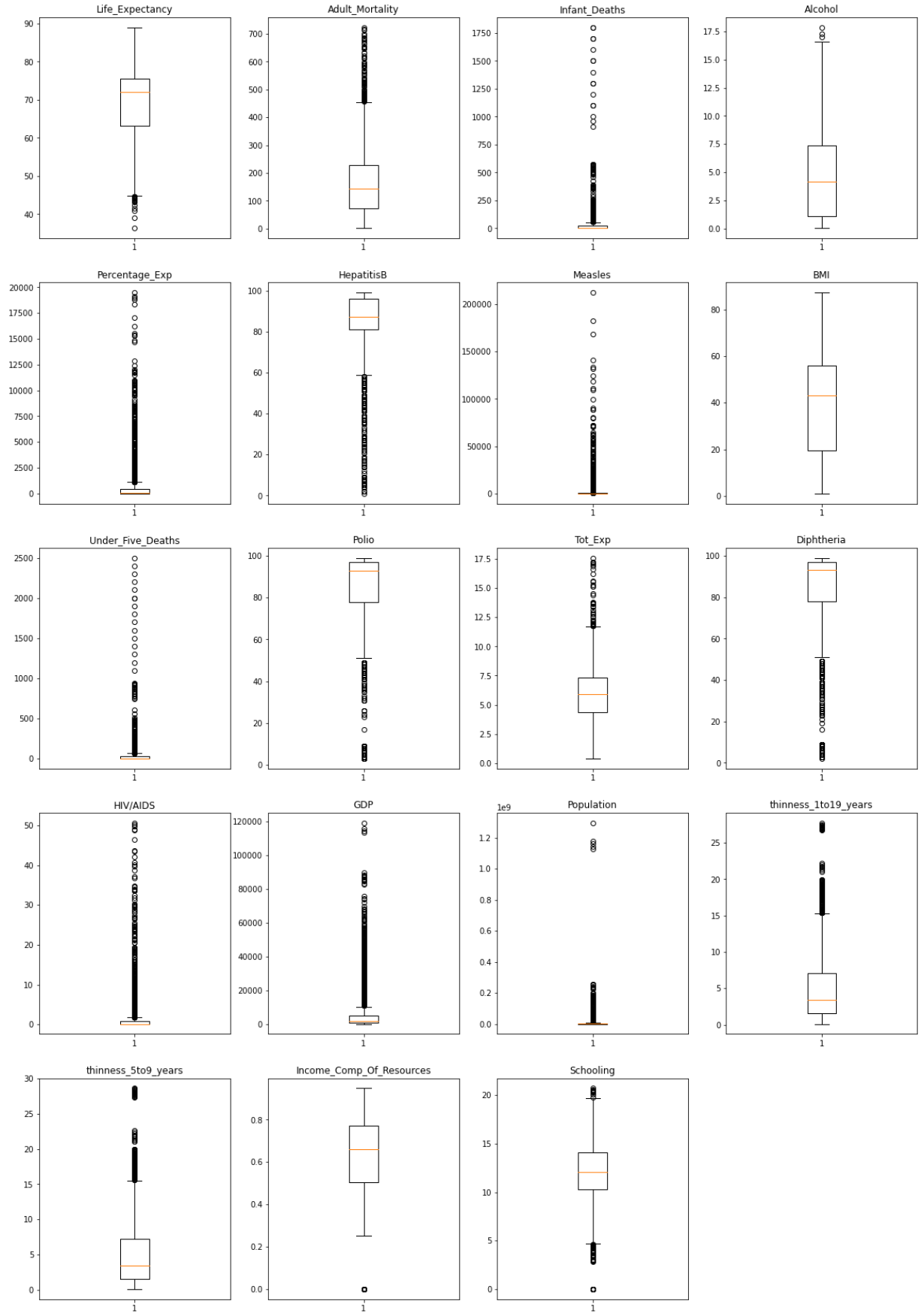
Out[57]:
```
Country                          0.000000
Year                             0.000000
Status                           0.000000
Life_Expectancy                  0.340368
Adult_Mortality                  0.340368
Infant_Deaths                    0.000000
Alcohol                          6.603131
Percentage_Exp                   0.000000
HepatitisB                      18.822328
Measles                          0.000000
BMI                              1.157250
Under_Five_Deaths                0.000000
Polio                            0.646698
Tot_Exp                          7.692308
Diphtheria                       0.646698
HIV/AIDS                         0.000000
GDP                             15.248468
Population                      22.191967
thinness_1to19_years             1.157250
thinness_5to9_years              1.157250
Income_Comp_Of_Resources         5.684139
Schooling                        5.547992
dtype: float64
```

In [58]:
```
#Impute values
df['Life_Expectancy'] = df['Life_Expectancy'].fillna(df['Life_Expectancy'].mean())
df['Adult_Mortality'] = df['Adult_Mortality'].fillna(df['Adult_Mortality'].mean())
df['Alcohol'] = df['Alcohol'].fillna(df['Alcohol'].mean())
df['HepatitisB'] = df['HepatitisB'].fillna(df['HepatitisB'].mean())
```

```
df['BMI'] = df['BMI'].fillna(df['BMI'].mean())
df['Polio'] = df['Polio'].fillna(df['Polio'].mean())
df['Tot_Exp'] = df['Tot_Exp'].fillna(df['Tot_Exp'].mean())
df['Diphtheria'] = df['Diphtheria'].fillna(df['Diphtheria'].mean())
df['GDP'] = df['GDP'].fillna(df['GDP'].median())
df['Population'] = df['Population'].fillna(df['Population'].median())
df['thinness_1to19_years'] = df['thinness_1to19_years'].fillna(df['thinness_1to19_ye
df['thinness_5to9_years'] = df['thinness_5to9_years'].fillna(df['thinness_5to9_years
df['Income_Comp_Of_Resources'] = df['Income_Comp_Of_Resources'].fillna(df['Income_Co
df['Schooling'] = df['Schooling'].fillna(df['Schooling'].mean())
```

In [59]:

```
# Create a dictionary of columns.
col_dict = {'Life_Expectancy':1,'Adult_Mortality':2,'Infant_Deaths':3,'Alcohol':4,'P


# Detect outliers in each variable using box plots.
plt.figure(figsize=(20,30))


for variable,i in col_dict.items():
                    plt.subplot(5,4,i)
                    plt.boxplot(df[variable],whis=1.5)
                    plt.title(variable)

plt.show()

#sns.boxplot(data = df, x='Population')
#plt.show()
```



In [60]:
```
df.columns
```

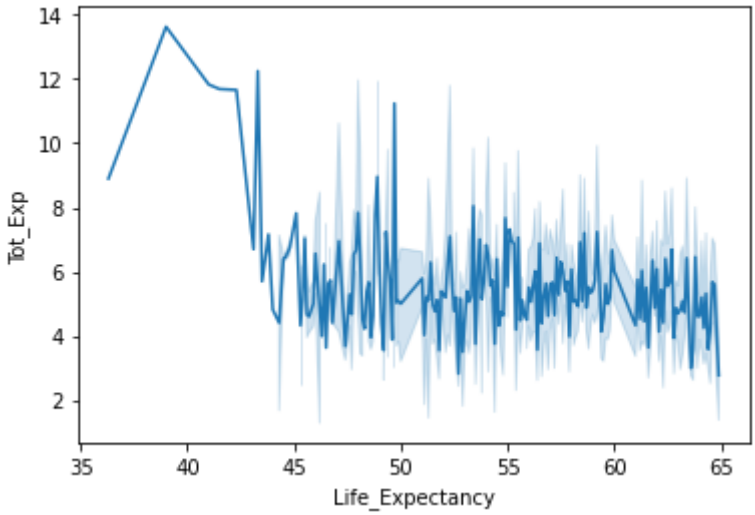
Out[60]:
```
Index(['Country', 'Year', 'Status', 'Life_Expectancy', 'Adult_Mortality',
       'Infant_Deaths', 'Alcohol', 'Percentage_Exp', 'HepatitisB', 'Measles',
       'BMI', 'Under_Five_Deaths', 'Polio', 'Tot_Exp', 'Diphtheria',
       'HIV/AIDS', 'GDP', 'Population', 'thinness_1to19_years',
       'thinness_5to9_years', 'Income_Comp_Of_Resources', 'Schooling'],
      dtype='object')
```

In [63]:
```
df.isnull().sum()
```

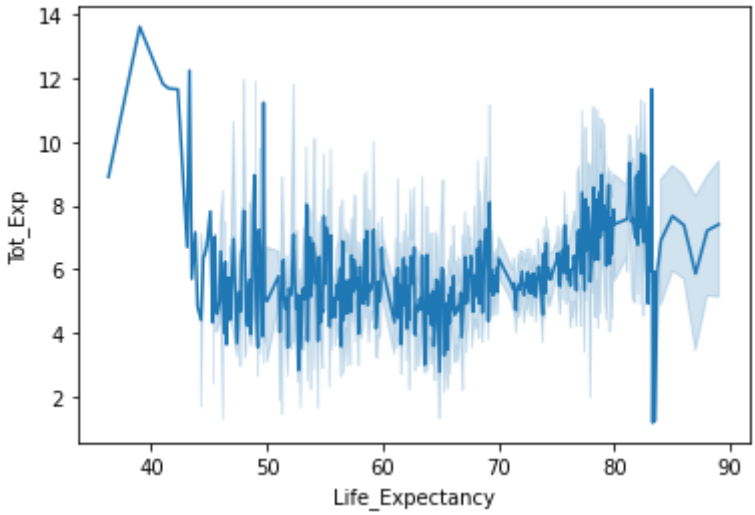
Out[63]:
```
Country                          0
Year                             0
Status                           0
Life_Expectancy                  0
Adult_Mortality                  0
Infant_Deaths                    0
Alcohol                          0
Percentage_Exp                   0
HepatitisB                       0
Measles                          0
BMI                              0
Under_Five_Deaths                0
Polio                            0
Tot_Exp                          0
Diphtheria                       0
HIV/AIDS                         0
GDP                              0
Population                       0
thinness_1to19_years             0
thinness_5to9_years              0
Income_Comp_Of_Resources         0
Schooling                        0
dtype: int64
```

In [66]:
```
#1. Should a country having a lower life expectancy value(<65) increase its healthca
# Target attribute is life expectancy

df_less65 = df[df["Life_Expectancy"] < 65]
```

In [69]:
```
sns.lineplot(x="Life_Expectancy", y="Tot_Exp", data = df_less65)
plt.show()
```

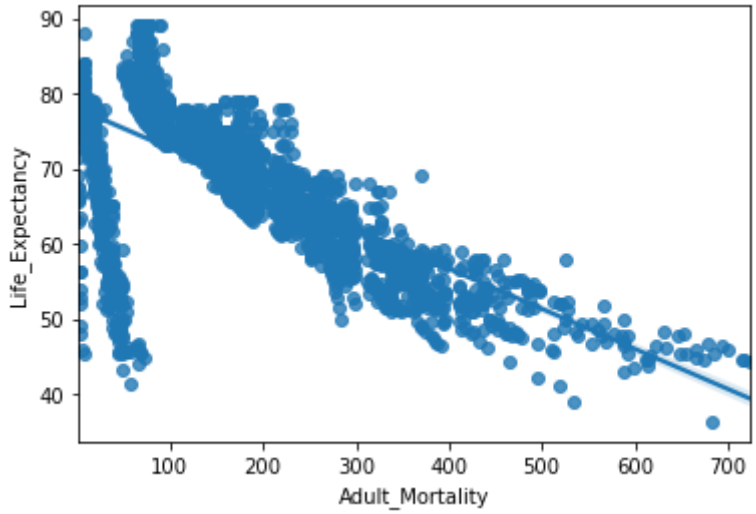


In [68]:
```
sns.lineplot(x="Life_Expectancy", y="Tot_Exp",data = df)
plt.show()
```

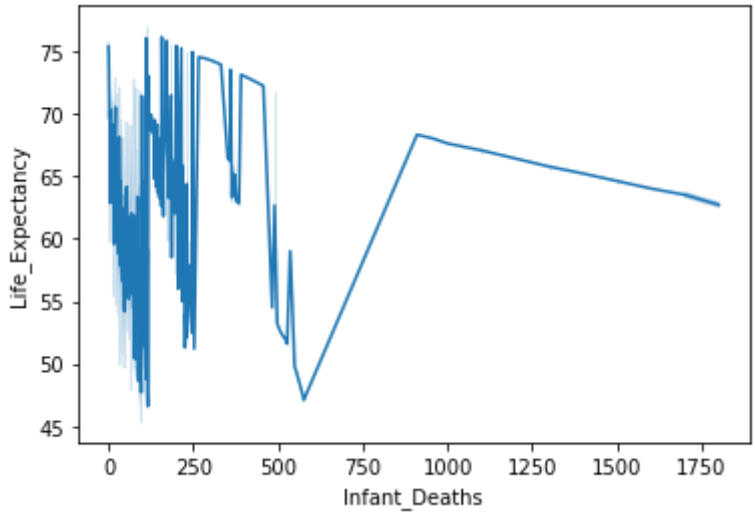


Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan? # Answer - No, The Increase in healthcare expenditure of the country does not improve its average life expectancy.

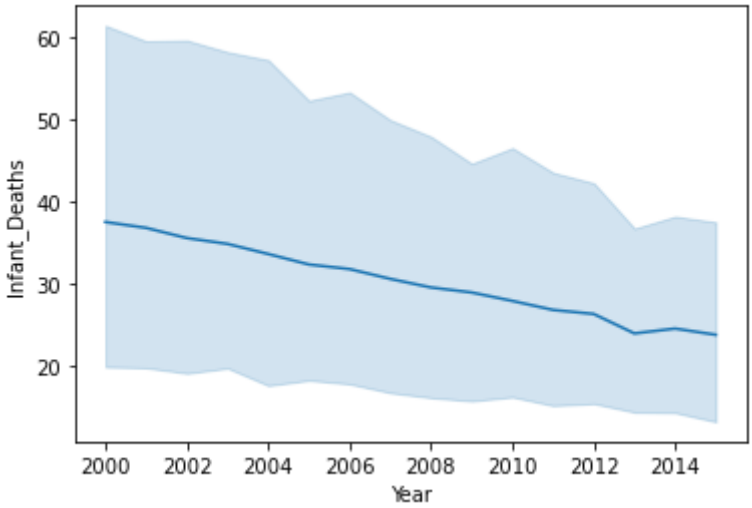
In [70]:
```
#2. How does Infant and Adult mortality rates affect life expectancy?

sns.regplot(x='Adult_Mortality',y='Life_Expectancy', data=df)
plt.show()
```



In [71]:
```
sns.lineplot(x='Infant_Deaths',y='Life_Expectancy',data=df)
plt.show()
#'infantdeaths'
```



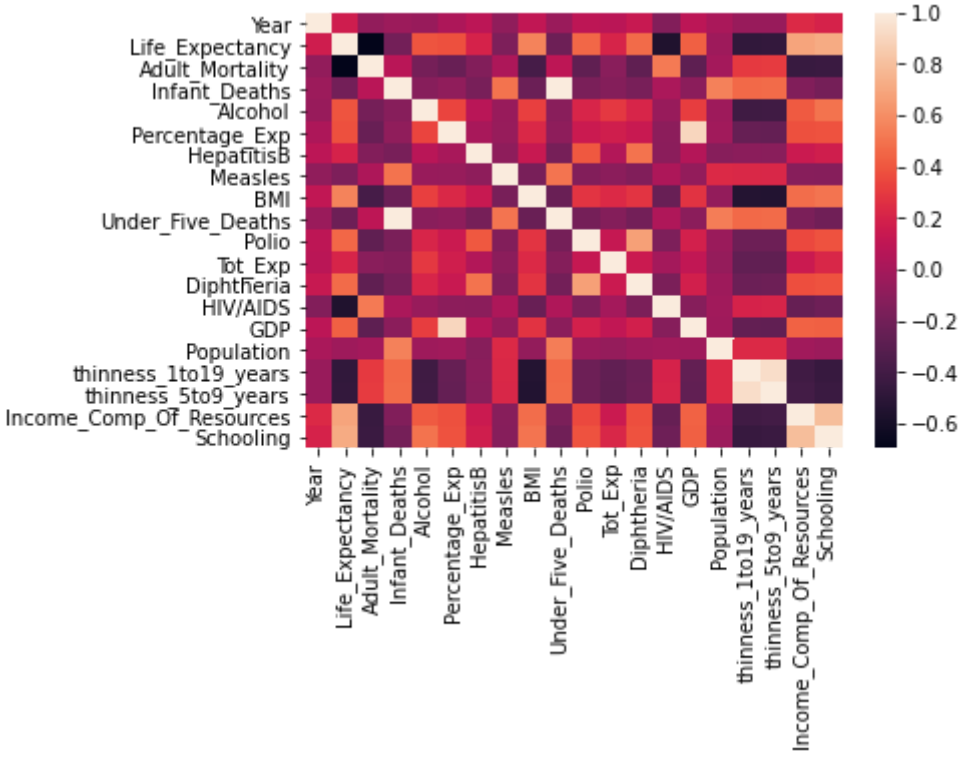
In [72]:
```
sns.lineplot(x='Year',y='Infant_Deaths',data=df)
plt.show()
#'infantdeaths'
```



#How does Infant and Adult mortality rates affect life expectancy? # Adult mortality rates are reducing when life expectancy is increasing # Infant mortality rates does not have strong corelation with life expectancy

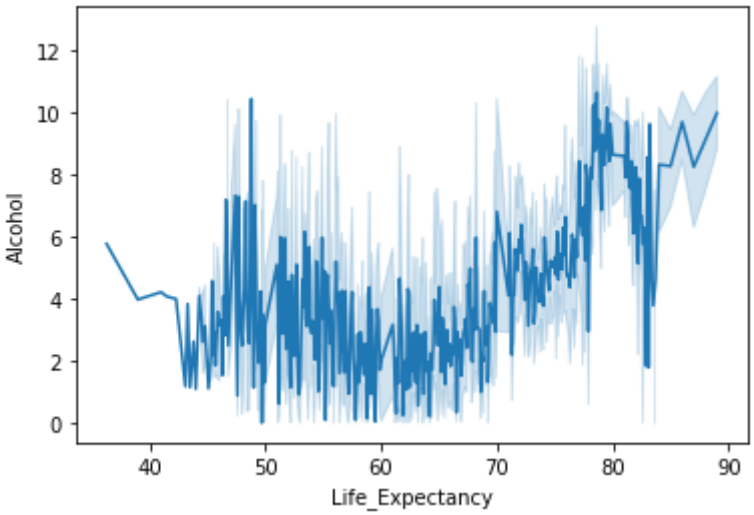
In [73]:
```
#3. Does Life Expectancy has positive or negative correlation with eating habits, li

corr = df.corr()
sns.heatmap(corr,xticklabels = corr.columns, yticklabels = corr.columns)
```

Out[73]: <AxesSubplot:>



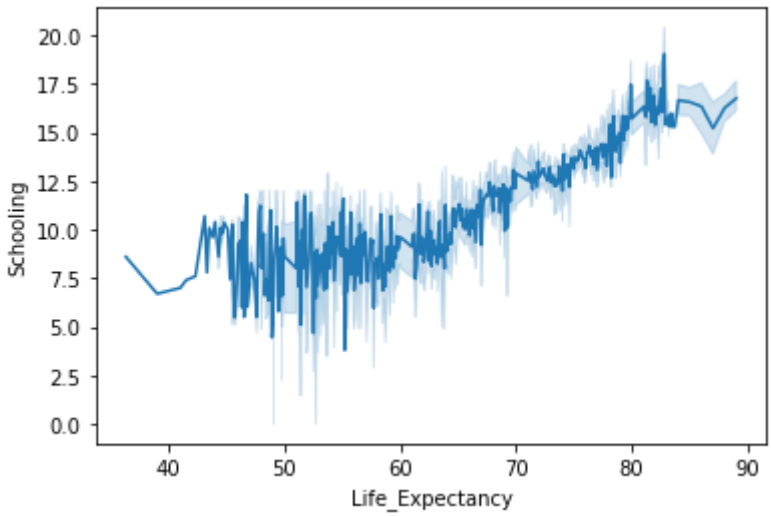
#Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc. # Alcohol - No correlation with Life Expectancy # BMI - No correlation with Life Expectancy

In [74]:
```
#4. Does Life Expectancy have positive or negative relationship with drinking alcoho

sns.lineplot(x='Life_Expectancy',y='Alcohol',data=df)
plt.show()
#'infantdeaths'
```



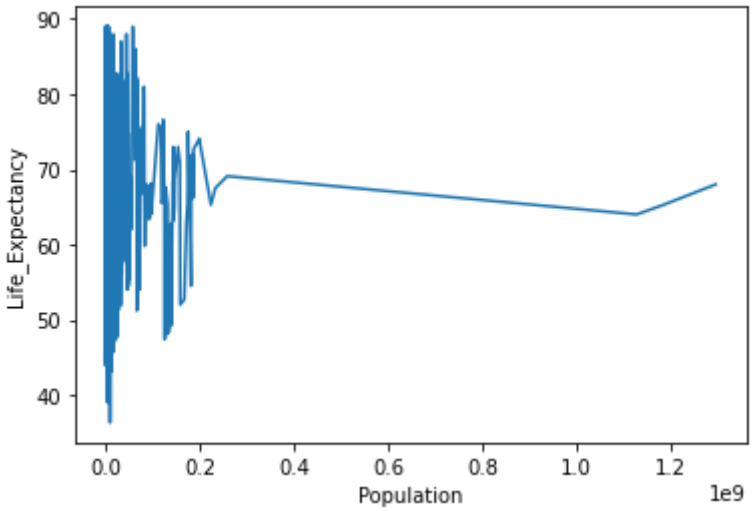
#Does Life Expectancy have positive or negative relationship with drinking alcohol? # No, Life Expectancy does not have positive or negative relationship with drinking alcohol

In [75]:
```
#5. What is the impact of schooling on the lifespan of humans?

sns.lineplot(x='Life_Expectancy',y='Schooling',data=df)
plt.show()
#'infantdeaths'
```

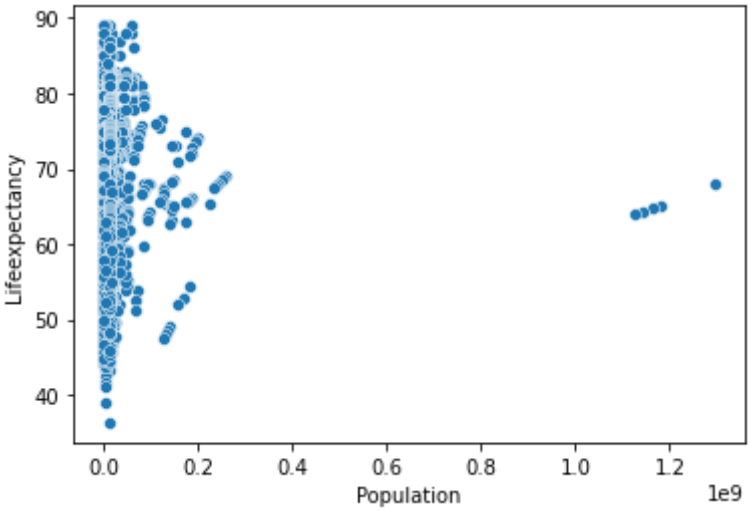


In [ ]:
```
#What is the impact of schooling on the lifespan of humans?

#Seeing postive impact of schooling against lifespan of humans
```

In [76]:
```
#6. Do densely populated countries tend to have lower life expectancy?
sns.lineplot(x='Population',y='Life_Expectancy',data=df)
plt.show()
#'infantdeaths'
```

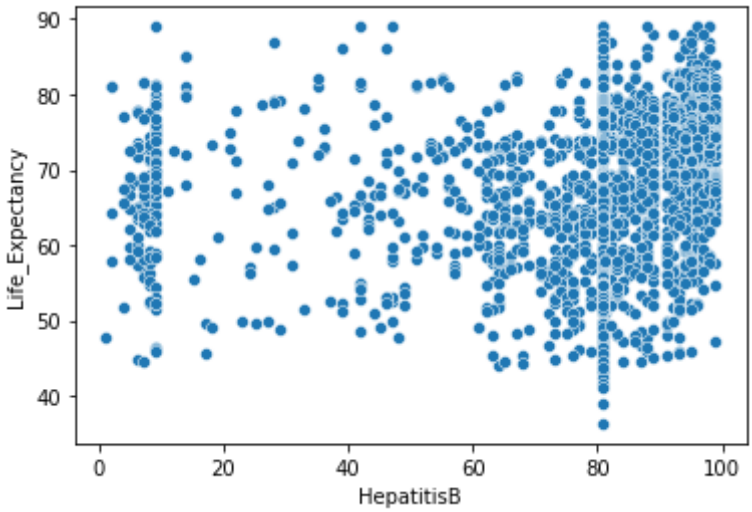


In [14]:
```
sns.scatterplot(x='Population',y='Life_Expectancy',data=df)
plt.show()
#'infantdeaths'
```

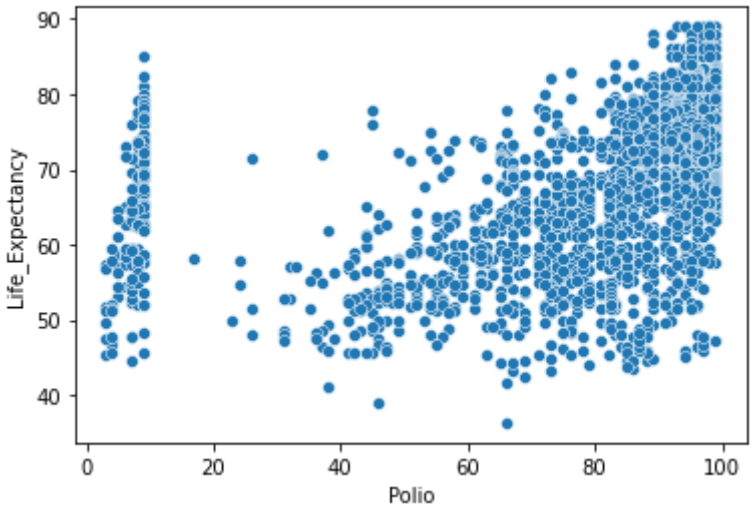


In [ ]:
```
#Do densely populated countries tend to have lower life expectancy?

# Densely populated countries have average life expectancy
```

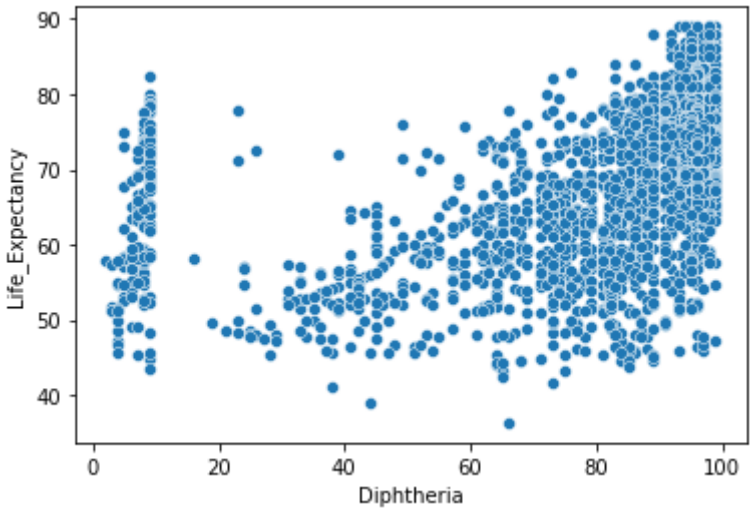
In [78]:
```
#7. What is the impact of Immunization coverage on life Expectancy?

sns.scatterplot(x='HepatitisB',y='Life_Expectancy',data=df)
plt.show()
```



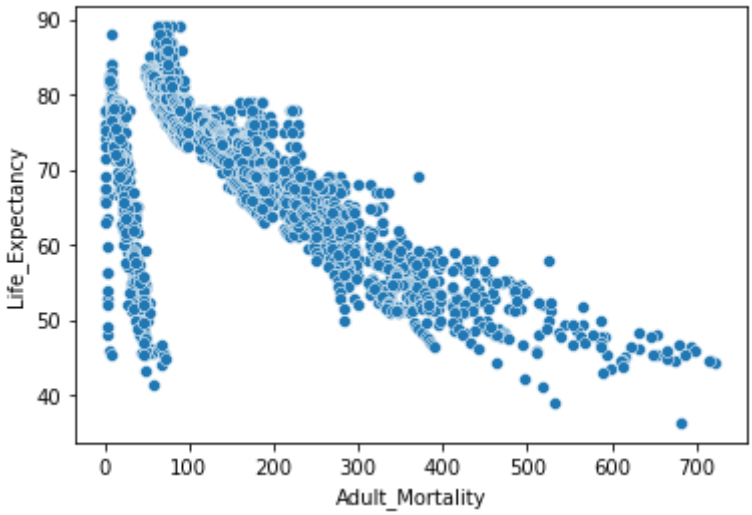
In [79]:
```
#What is the impact of Immunization coverage on life Expectancy?

sns.scatterplot(x='Polio',y='Life_Expectancy',data=df)
plt.show()
```



In [80]:
```
#What is the impact of Immunization coverage on life Expectancy?

sns.scatterplot(x='Diphtheria',y='Life_Expectancy',data=df)
plt.show()
```



In [81]:
```
#What is the impact of Immunization coverage on life Expectancy?

sns.scatterplot(x='Adult_Mortality',y='Life_Expectancy',data=df)
plt.show()
```



In [ ]:
```
# Diphtheria and Polio are correlated. We can drop one of the column.
# Adult mortality has negative relationship  with Life expectancy
```