---------- The data analysis focuses on a dataset related to health, specifically a case study of life expectancy in Afghanistan. This dataset is sourced from Kaggle.----------

----------Import Libraries----------

In [44]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

----------Read Dataset----------

In [45]:
```
data = pd.read_csv('C:/Users/DELL/Desktop/Atomcamp Python/Life Expectancy Data.csv'
```

In [46]:
```
data.head(10)
```

Out[46]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measle
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	115
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	49
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	43
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	278
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	301
5	Afghanistan	2010	Developing	58.8	279.0	74	0.01	79.679367	66.0	198
6	Afghanistan	2009	Developing	58.6	281.0	77	0.01	56.762217	63.0	286
7	Afghanistan	2008	Developing	58.1	287.0	80	0.03	25.873925	64.0	159
8	Afghanistan	2007	Developing	57.5	295.0	82	0.02	10.910156	63.0	114
9	Afghanistan	2006	Developing	57.3	295.0	84	0.03	17.171518	64.0	199

10 rows × 22 columns

In [48]:
```
data.describe()
```

Out[48]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000

In [49]:
```
data.isnull().sum()
```

Out[49]:
```
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
```

In [50]:
```
data.dropna(inplace=True)
```

In [52]:
```
data.isnull().sum()
```

Out[52]:
```
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
```

In [53]:
```
data.head()
```

Out[53]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measle
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	115
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	49
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	43
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	278
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	301

5 rows × 22 columns

In [54]:
```
data.tail()
```

Out[54]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Meas
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.0	68.0	
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.0	7.0	9
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.0	73.0	1
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.0	76.0	5
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.0	79.0	14

5 rows × 22 columns

----------Data Sanity Check----------

In [55]:
```
data.shape # Clean data without missing and duplicate values
```

Out[55]:
```
(1649, 22)
```

In [56]:
```
data.info() # Cleaned data without any error
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1649 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          1649 non-null   object 
 1   Year                             1649 non-null   int64  
 2   Status                           1649 non-null   object 
 3   Life expectancy                  1649 non-null   float64
 4   Adult Mortality                  1649 non-null   float64
 5   infant deaths                    1649 non-null   int64  
 6   Alcohol                          1649 non-null   float64
 7   percentage expenditure           1649 non-null   float64
 8   Hepatitis B                      1649 non-null   float64
 9   Measles                          1649 non-null   int64  
 10   BMI                             1649 non-null   float64
 11  under-five deaths                1649 non-null   int64  
 12  Polio                            1649 non-null   float64
 13  Total expenditure                1649 non-null   float64
 14  Diphtheria                       1649 non-null   float64
 15   HIV/AIDS                        1649 non-null   float64
 16  GDP                              1649 non-null   float64
 17  Population                       1649 non-null   float64
 18   thinness  1-19 years            1649 non-null   float64
 19   thinness 5-9 years              1649 non-null   float64
 20  Income composition of resources  1649 non-null   float64
 21  Schooling                        1649 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 296.3+ KB
```

In [57]:
```
df.duplicated().sum()
```

Out[57]:
```
0
```

In [63]:
```
for i in df.select_dtypes(include='object').columns:
    print(df[i].value_counts())
    print("-----"*10)
```

```
$53,594.10       1
$18,421.20       1
Name: count, Length: 559, dtype: int64
--------------------------------------------------
COGS
$17,430.00     4
$8,655.00      3
$24,700.00     3
$1,101.00      3
$15,140.00     3
              ..
$20,300.00     1
$1,686.00      1
$11,175.00     1
$38,010.00     1
$5,418.00      1
Name: count, Length: 545, dtype: int64
--------------------------------------------------
Profit
$-             5
```

----------Exploratory Data Analysis ----------

In [65]:
```
data.describe().T
```

Out[65]:

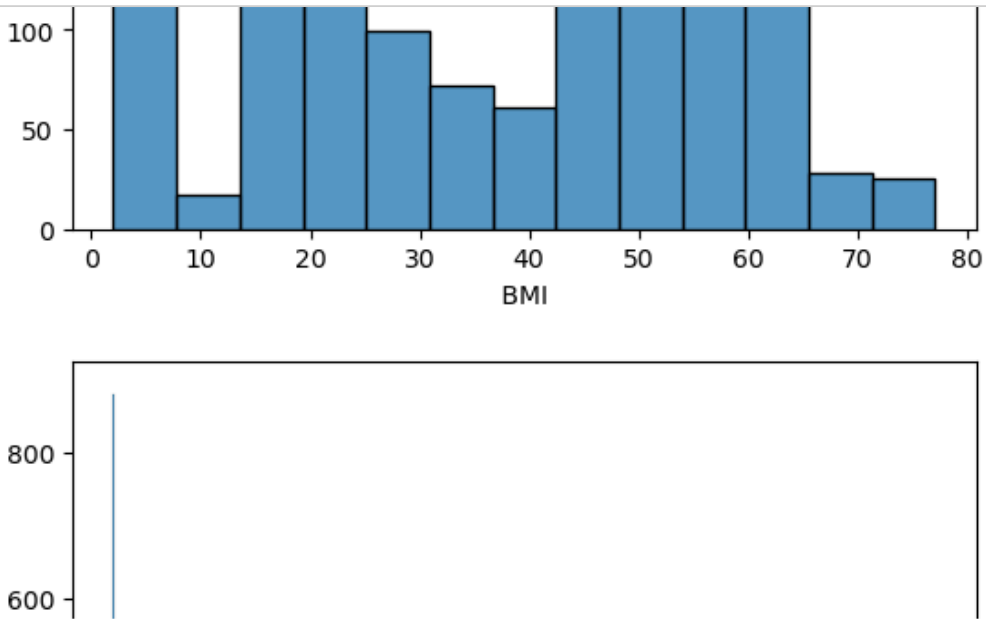
	count	mean	std	min	25%	50%	75
Year	1649.0	2.007841e+03	4.087711e+00	2000.00000	2005.000000	2.008000e+03	2.011000e+
Life expectancy	1649.0	6.930230e+01	8.796834e+00	44.00000	64.400000	7.170000e+01	7.500000e+
Adult Mortality	1649.0	1.682153e+02	1.253104e+02	1.00000	77.000000	1.480000e+02	2.270000e+
infant deaths	1649.0	3.255306e+01	1.208472e+02	0.00000	1.000000	3.000000e+00	2.200000e+
Alcohol	1649.0	4.533196e+00	4.029189e+00	0.01000	0.810000	3.790000e+00	7.340000e+
percentage expenditure	1649.0	6.989736e+02	1.759229e+03	0.00000	37.438577	1.451023e+02	5.093900e+
Hepatitis B	1649.0	7.921771e+01	2.560466e+01	2.00000	74.000000	8.900000e+01	9.600000e+
Measles	1649.0	2.224494e+03	1.008580e+04	0.00000	0.000000	1.500000e+01	3.730000e+
BMI	1649.0	3.812862e+01	1.975425e+01	2.00000	19.500000	4.370000e+01	5.580000e+
under-five deaths	1649.0	4.422013e+01	1.628980e+02	0.00000	1.000000	4.000000e+00	2.900000e+
Polio	1649.0	8.356458e+01	2.245056e+01	3.00000	81.000000	9.300000e+01	9.700000e+
Total expenditure	1649.0	5.955925e+00	2.299385e+00	0.74000	4.410000	5.840000e+00	7.470000e+
Diphtheria	1649.0	8.415525e+01	2.157919e+01	2.00000	82.000000	9.200000e+01	9.700000e+
HIV/AIDS	1649.0	1.983869e+00	6.032360e+00	0.10000	0.100000	1.000000e-01	7.000000e-
GDP	1649.0	5.566032e+03	1.147590e+04	1.68135	462.149650	1.592572e+03	4.718513e+
Population	1649.0	1.465363e+07	7.046039e+07	34.00000	191897.000000	1.419631e+06	7.658972e+
thinness 1-19 years	1649.0	4.850637e+00	4.599228e+00	0.10000	1.600000	3.000000e+00	7.100000e+
thinness 5-9 years	1649.0	4.907762e+00	4.653757e+00	0.10000	1.700000	3.200000e+00	7.100000e+
Income composition of resources	1649.0	6.315512e-01	1.830887e-01	0.00000	0.509000	6.730000e-01	7.510000e-
Schooling	1649.0	1.211989e+01	2.795388e+00	4.20000	10.300000	1.230000e+01	1.400000e+

In [67]:
```
data.describe(include='object').T
```

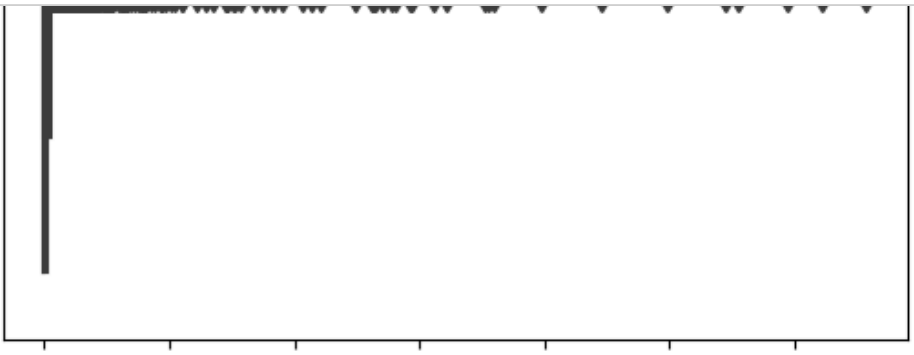
Out[67]:

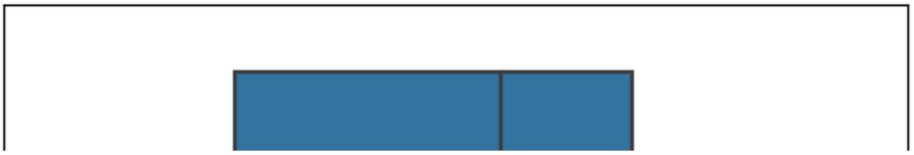
	count	unique	top	freq
Country	1649	133	Afghanistan	16
Status	1649	2	Developing	1407

In [69]:
```
for i in data.select_dtypes(include='number').columns:
    sns.histplot(data=data,x=i)
    plt.show()
```



In [70]:
```
for i in data.select_dtypes(include='number').columns:
    sns.boxplot(data=data,x=i)
    plt.show()
```



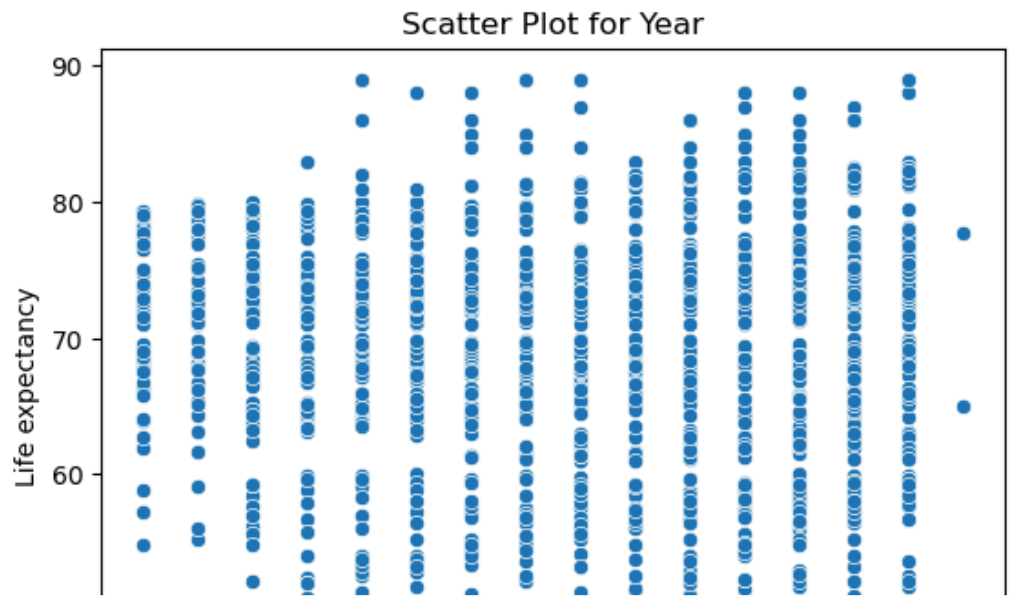


In [71]:
```
data.select_dtypes(include='number').columns
```

Out[71]:
```
Index(['Year', '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'],
      dtype='object')
```

In [83]:
```
cols = ['Year', 'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expendit
        '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 col in cols:
    sns.scatterplot(data=data, x=col, y='Life expectancy ')
    plt.title(f'Scatter Plot for {col}')
    plt.xlabel(col)
    plt.ylabel('Life expectancy')
    plt.show()
```



In [84]:
```
corelation = data.select_dtypes(include='number').corr()
```

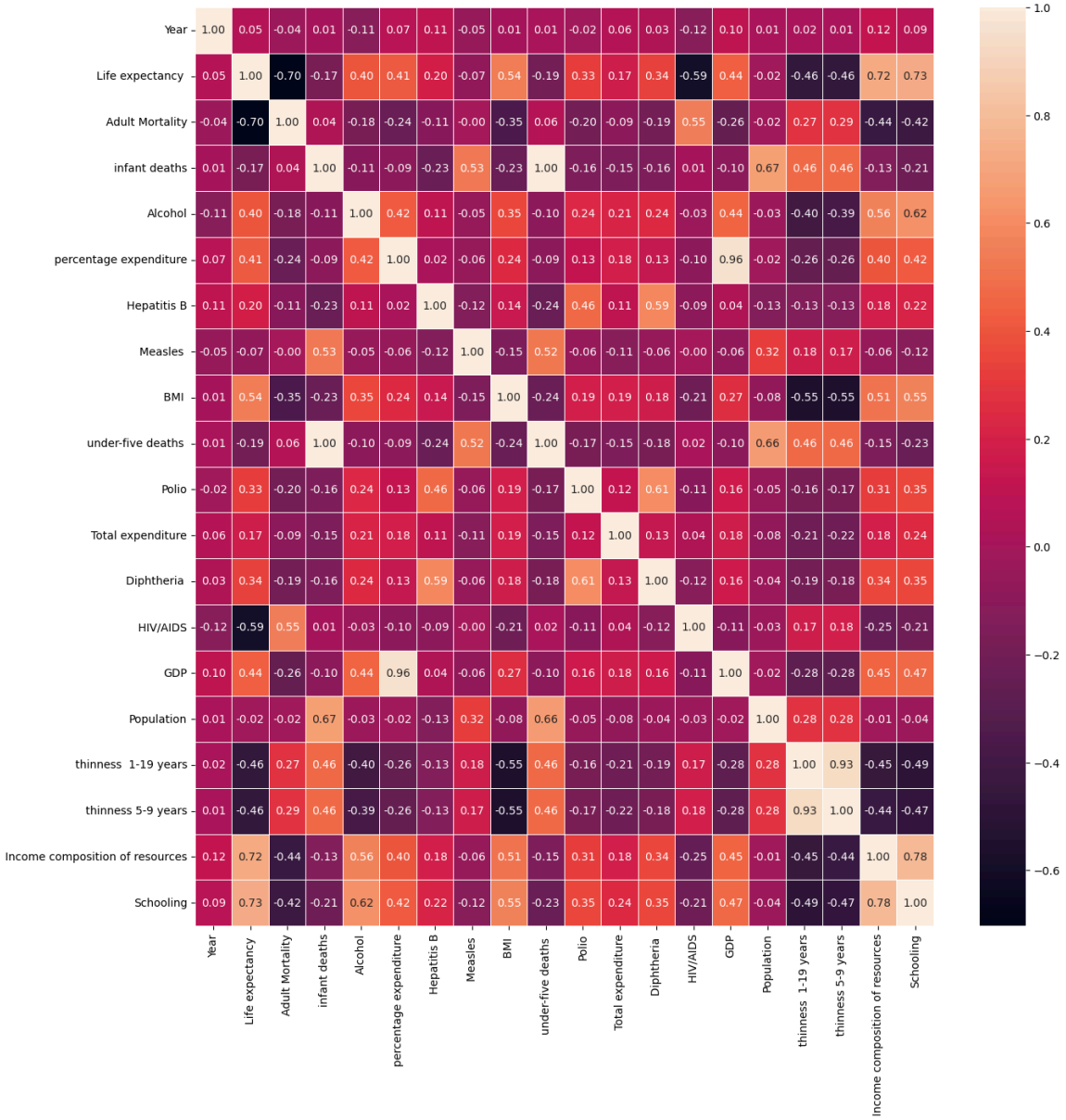
In [85]:
```
corelation
```

Out[85]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles
Year	1.000000	0.050771	-0.037092	0.008029	-0.113365	0.069553	0.114897	-0.053822
Life expectancy	0.050771	1.000000	-0.702523	-0.169074	0.402718	0.409631	0.199935	-0.068881
Adult Mortality	-0.037092	-0.702523	1.000000	0.042450	-0.175535	-0.237610	-0.105225	-0.003967
infant deaths	0.008029	-0.169074	0.042450	1.000000	-0.106217	-0.090765	-0.231769	0.532680
Alcohol	-0.113365	0.402718	-0.175535	-0.106217	1.000000	0.417047	0.109889	-0.050110
percentage expenditure	0.069553	0.409631	-0.237610	-0.090765	0.417047	1.000000	0.016760	-0.063071
Hepatitis B	0.114897	0.199935	-0.105225	-0.231769	0.109889	0.016760	1.000000	-0.124800
Measles	-0.053822	-0.068881	-0.003967	0.532680	-0.050110	-0.063071	-0.124800	1.000000
BMI	0.005739	0.542042	-0.351542	-0.234425	0.353396	0.242738	0.143302	-0.153245
under-five deaths	0.010479	-0.192265	0.060365	0.996906	-0.101082	-0.092158	-0.240766	0.517506
Polio	-0.016699	0.327294	-0.199853	-0.156929	0.240315	0.128626	0.463331	-0.057850
Total expenditure	0.059493	0.174718	-0.085227	-0.146951	0.214885	0.183872	0.113327	-0.113583
Diphtheria	0.029641	0.341331	-0.191429	-0.161871	0.242951	0.134813	0.588990	-0.058606
HIV/AIDS	-0.123405	-0.592236	0.550691	0.007712	-0.027113	-0.095085	-0.094802	-0.003522
GDP	0.096421	0.441322	-0.255035	-0.098092	0.443433	0.959299	0.041850	-0.064768
Population	0.012567	-0.022305	-0.015012	0.671758	-0.028880	-0.016792	-0.129723	0.321946
thinness 1-19 years	0.019757	-0.457838	0.272230	0.463415	-0.403755	-0.255035	-0.129406	0.180642
thinness 5-9 years	0.014122	-0.457508	0.286723	0.461908	-0.386208	-0.255635	-0.133251	0.174946
Income composition of resources	0.122892	0.721083	-0.442203	-0.134754	0.561074	0.402170	0.184921	-0.058277
Schooling	0.088732	0.727630	-0.421171	-0.214372	0.616975	0.422088	0.215182	-0.115660

In [86]:
```
plt.figure(figsize=(15,15))
sns.heatmap(corelation, annot=True, fmt=".2f",linewidths=.5)
```

Out[86]: <Axes: >



Outliers Treatment

In [100]:
```
def wisker(cpl):
    q1, q3 = np.percentile(cpl, (25, 75))
    iqr = q3 - q1
    lw = q1 - 1.5 * iqr
    uw = q3 + 1.5 * iqr
    return lw, uw
```

In [101]:
```
wisker(data[ 'Life expectancy '])
```

Out[101]: (48.500000000000014, 90.89999999999999)

----------THE END ----------

In [ ]: