Life Expectancy data analysis

August 16, 2024

1 Data preprocessing before building the machine learning model

1.1 STEPS of preprocessing data

#step:1] import the necessary library #step:2] Read dataset #step:3] sanity check of data #step:4] exploratory data analysis(EDA) #step:5] missing value treatment #step:6] outlier treatment #step:7] duplicate garbag value treatment #step:8] noemalization #step:9] encoding od data

[32]:
```
##Import necessary library
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

[33]:
```
##Read the data
data = pd.read_csv("Life Expectancy Data.csv")
```

[34]:
```
## find the top 3 rows
data.head(3)
```

[34]:

	Country	Year	Status	Life expectancy	Adult Mortality	\
0	Afghanistan	2015	Developing	65.0	263.0	
1	Afghanistan	2014	Developing	59.9	271.0	
2	Afghanistan	2013	Developing	59.9	268.0	

	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	…	\
0	62	0.01	71.279624	65.0	1154	…	
1	64	0.01	73.523582	62.0	492	…	
2	66	0.01	73.219243	64.0	430	…	

	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	\
0	6.0	8.16	65.0	0.1	584.259210	33736494.0	
1	58.0	8.18	62.0	0.1	612.696514	327582.0	
2	62.0	8.13	64.0	0.1	631.744976	31731688.0	

	thinness 1-19 years	thinness 5-9 years	\
0	17.2	17.3	
1	17.5	17.5	

```
2                   17.7                 17.7

   Income composition of resources  Schooling
0                            0.479       10.1
1                            0.476       10.0
2                            0.470        9.9

[3 rows x 22 columns]
```

[35]:
```
##sanity check the data
data.shape
```

[35]:
```
(2938, 22)
```

[36]:
```
data.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
```

[37]:
```
data.isnull().sum()
```

[37]:
```
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
```

[38]:
```
data.isnull().sum()/data.shape[0]*100
```

[38]:
```
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 expenditure             0.000000
Hepatitis B                       18.822328
Measles                            0.000000
 BMI                               1.157250
under-five deaths                  0.000000
Polio                              0.646698
Total expenditure                  7.692308
Diphtheria                         0.646698
 HIV/AIDS                          0.000000
GDP                               15.248468
Population                        22.191967
 thinness  1-19 years              1.157250
 thinness 5-9 years                1.157250
Income composition of resources    5.684139
```

```
Schooling                           5.547992
dtype: float64
```

[39]:
```
data.duplicated().sum()
```

[39]:
```
0
```

[40]:
```
##EDA
data.describe().T
```

[40]:

	count	mean	std	\
Year	2938.0	2.007519e+03	4.613841e+00	
Life expectancy	2928.0	6.922493e+01	9.523867e+00	
Adult Mortality	2928.0	1.647964e+02	1.242921e+02	
infant deaths	2938.0	3.030395e+01	1.179265e+02	
Alcohol	2744.0	4.602861e+00	4.052413e+00	
percentage expenditure	2938.0	7.382513e+02	1.987915e+03	
Hepatitis B	2385.0	8.094046e+01	2.507002e+01	
Measles	2938.0	2.419592e+03	1.146727e+04	
BMI	2904.0	3.832125e+01	2.004403e+01	
under-five deaths	2938.0	4.203574e+01	1.604455e+02	
Polio	2919.0	8.255019e+01	2.342805e+01	
Total expenditure	2712.0	5.938190e+00	2.498320e+00	
Diphtheria	2919.0	8.232408e+01	2.371691e+01	
HIV/AIDS	2938.0	1.742103e+00	5.077785e+00	
GDP	2490.0	7.483158e+03	1.427017e+04	
Population	2286.0	1.275338e+07	6.101210e+07	
thinness 1-19 years	2904.0	4.839704e+00	4.420195e+00	
thinness 5-9 years	2904.0	4.870317e+00	4.508882e+00	
Income composition of resources	2771.0	6.275511e-01	2.109036e-01	
Schooling	2775.0	1.199279e+01	3.358920e+00	

	min	25%	50%	\
Year	2000.00000	2004.000000	2.008000e+03	
Life expectancy	36.30000	63.100000	7.210000e+01	
Adult Mortality	1.00000	74.000000	1.440000e+02	
infant deaths	0.00000	0.000000	3.000000e+00	
Alcohol	0.01000	0.877500	3.755000e+00	
percentage expenditure	0.00000	4.685343	6.491291e+01	
Hepatitis B	1.00000	77.000000	9.200000e+01	
Measles	0.00000	0.000000	1.700000e+01	
BMI	1.00000	19.300000	4.350000e+01	
under-five deaths	0.00000	0.000000	4.000000e+00	
Polio	3.00000	78.000000	9.300000e+01	
Total expenditure	0.37000	4.260000	5.755000e+00	
Diphtheria	2.00000	78.000000	9.300000e+01	
HIV/AIDS	0.10000	0.100000	1.000000e-01	

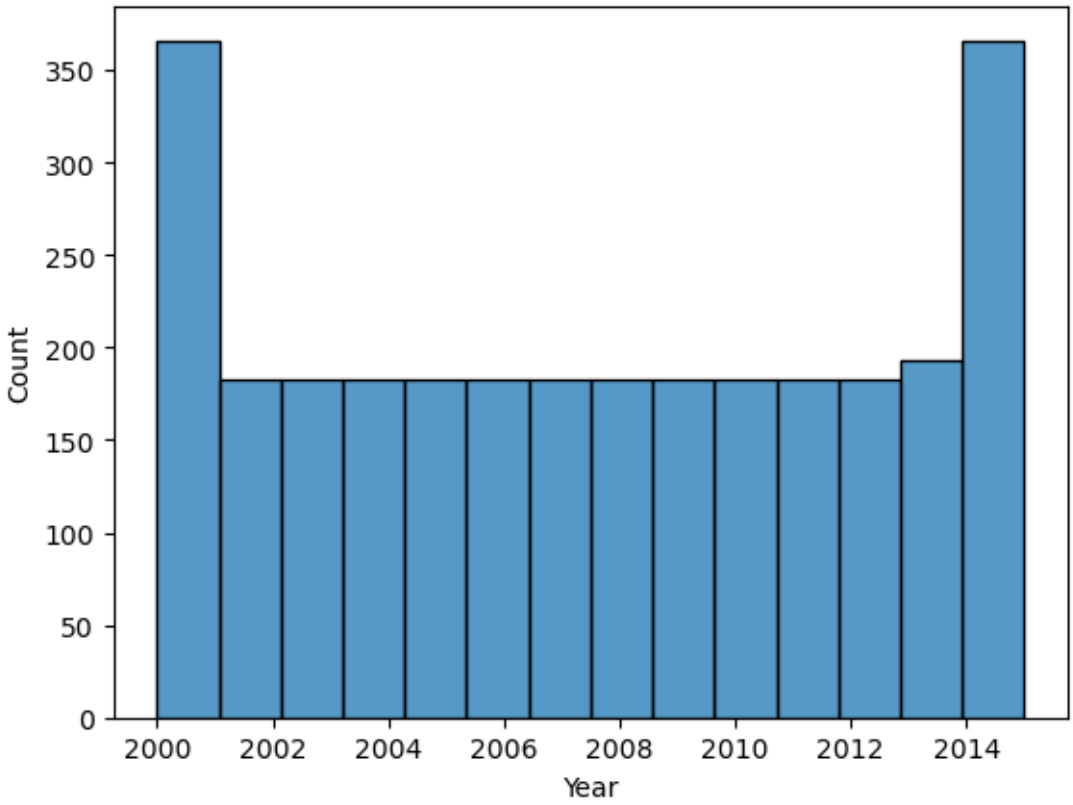
```
GDP                                1.68135     463.935626  1.766948e+03
Population                        34.00000  195793.250000  1.386542e+06
 thinness  1-19 years              0.10000       1.600000  3.300000e+00
 thinness 5-9 years               0.10000       1.500000  3.300000e+00
Income composition of resources     0.00000       0.493000  6.770000e-01
Schooling                           0.00000      10.100000  1.230000e+01

                                         75%           max
Year                            2.012000e+03  2.015000e+03
Life expectancy                 7.570000e+01  8.900000e+01
Adult Mortality                 2.280000e+02  7.230000e+02
infant deaths                   2.200000e+01  1.800000e+03
Alcohol                         7.702500e+00  1.787000e+01
percentage expenditure          4.415341e+02  1.947991e+04
Hepatitis B                     9.700000e+01  9.900000e+01
Measles                         3.602500e+02  2.121830e+05
 BMI                            5.620000e+01  8.730000e+01
under-five deaths               2.800000e+01  2.500000e+03
Polio                           9.700000e+01  9.900000e+01
Total expenditure               7.492500e+00  1.760000e+01
Diphtheria                      9.700000e+01  9.900000e+01
 HIV/AIDS                       8.000000e-01  5.060000e+01
GDP                             5.910806e+03  1.191727e+05
Population                      7.420359e+06  1.293859e+09
 thinness  1-19 years           7.200000e+00  2.770000e+01
 thinness 5-9 years             7.200000e+00  2.860000e+01
Income composition of resources 7.790000e-01  9.480000e-01
Schooling                       1.430000e+01  2.070000e+01
```

[41]:
```
data.describe(include="object")
```

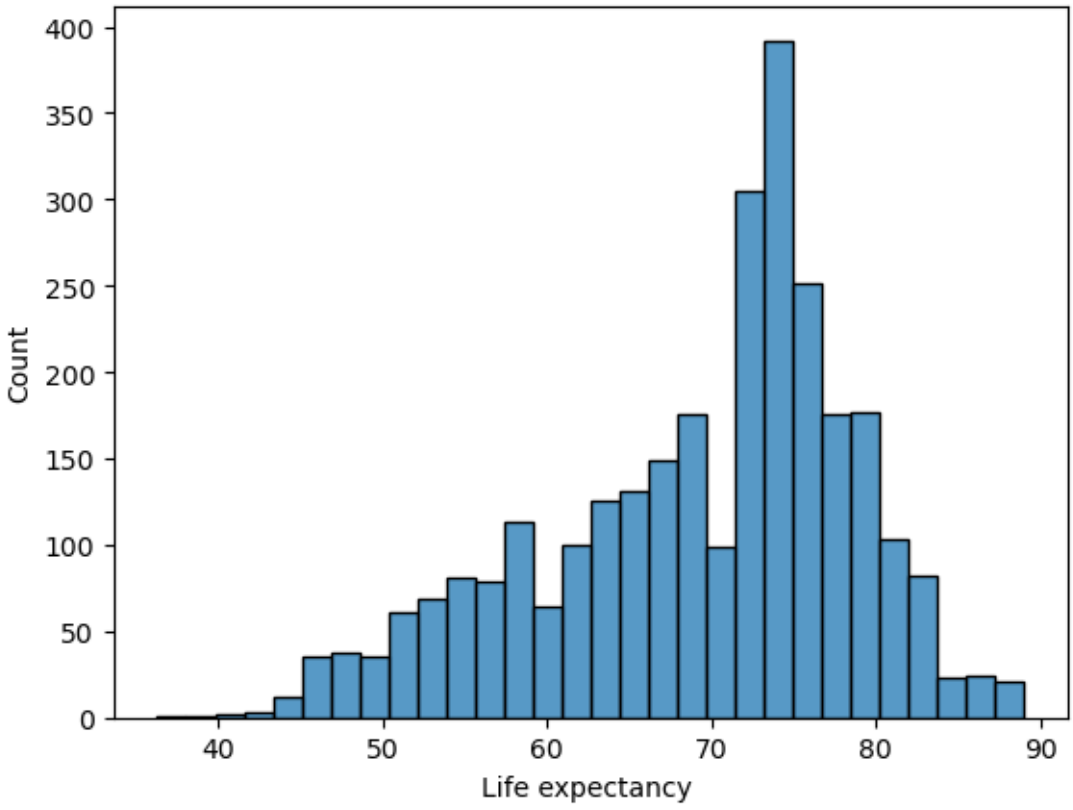
[41]:
```
            Country      Status
count          2938        2938
unique          193           2
top     Afghanistan  Developing
freq             16        2426
```

[42]:
```
## histogram to understand the destribution
for i in data.select_dtypes(include="number").columns:
    sns.histplot(data=data, x=i)  # Corrected data argument
    plt.show()  # Correctly close the function call
```

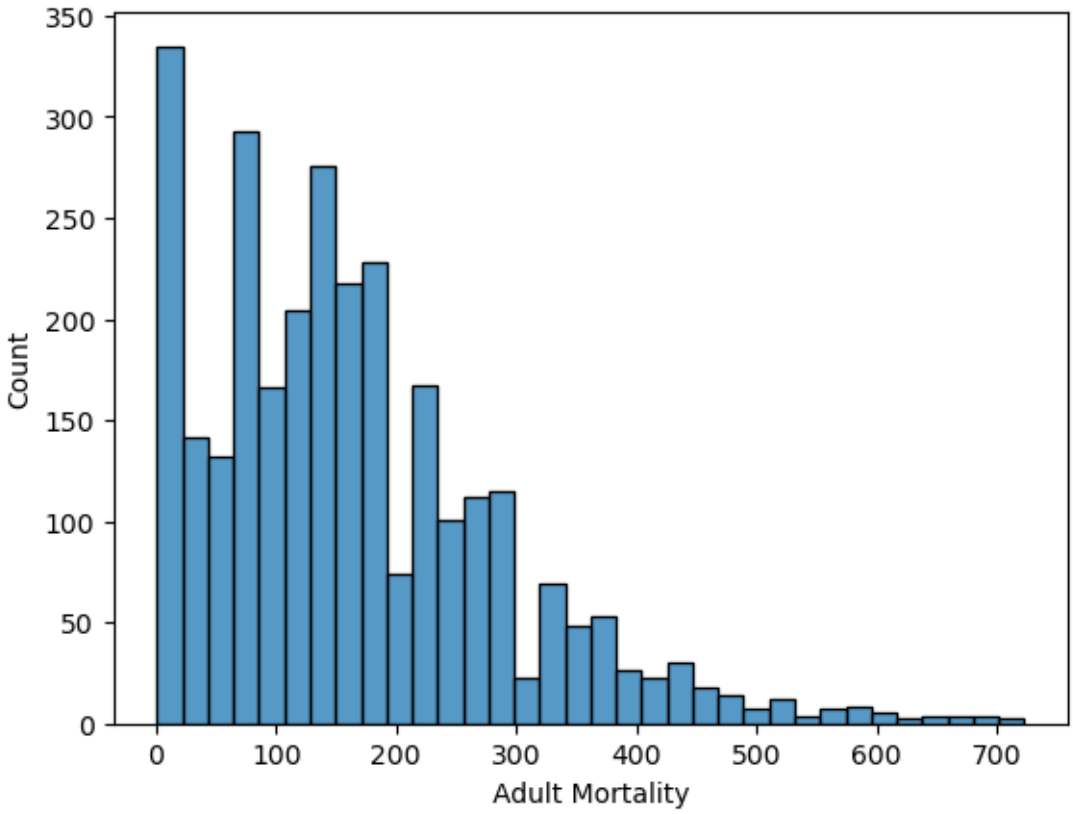
```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



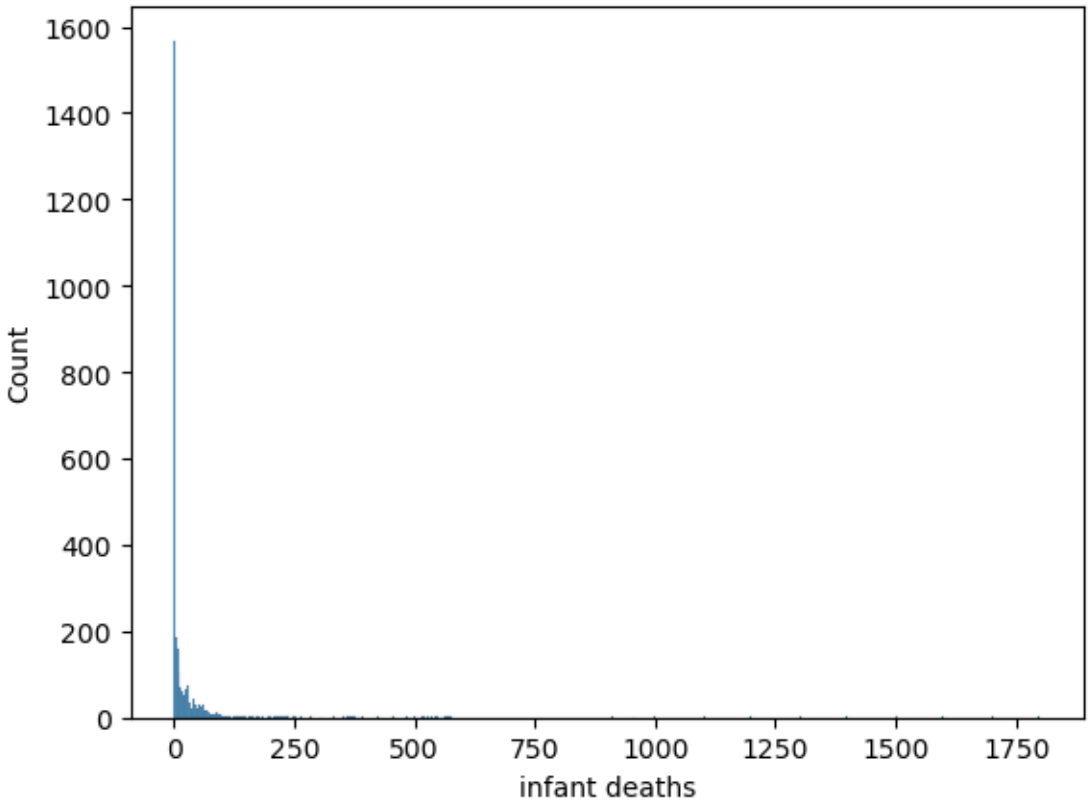

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



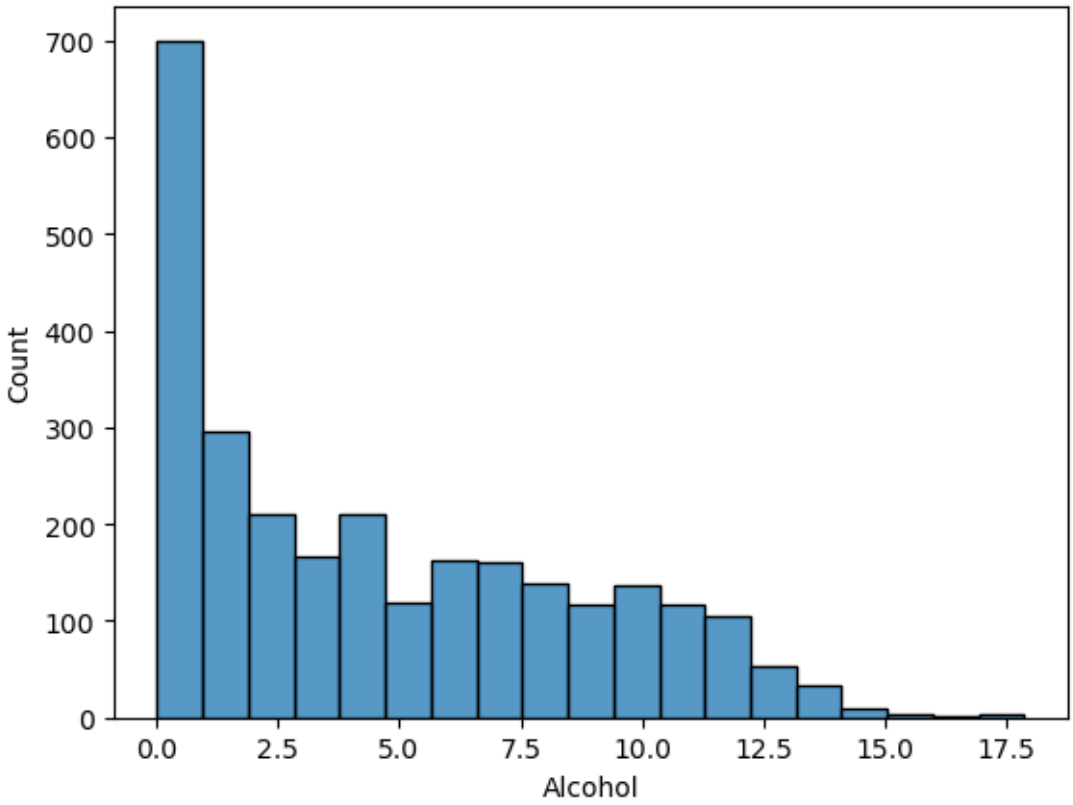

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



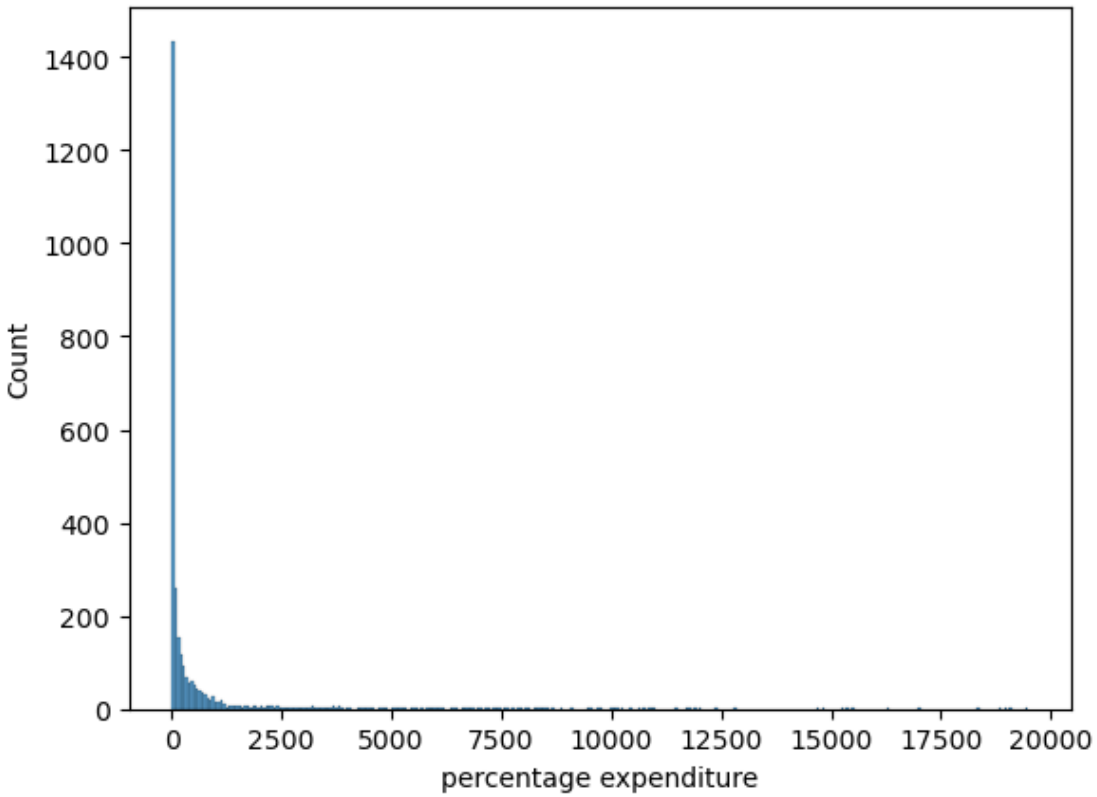

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



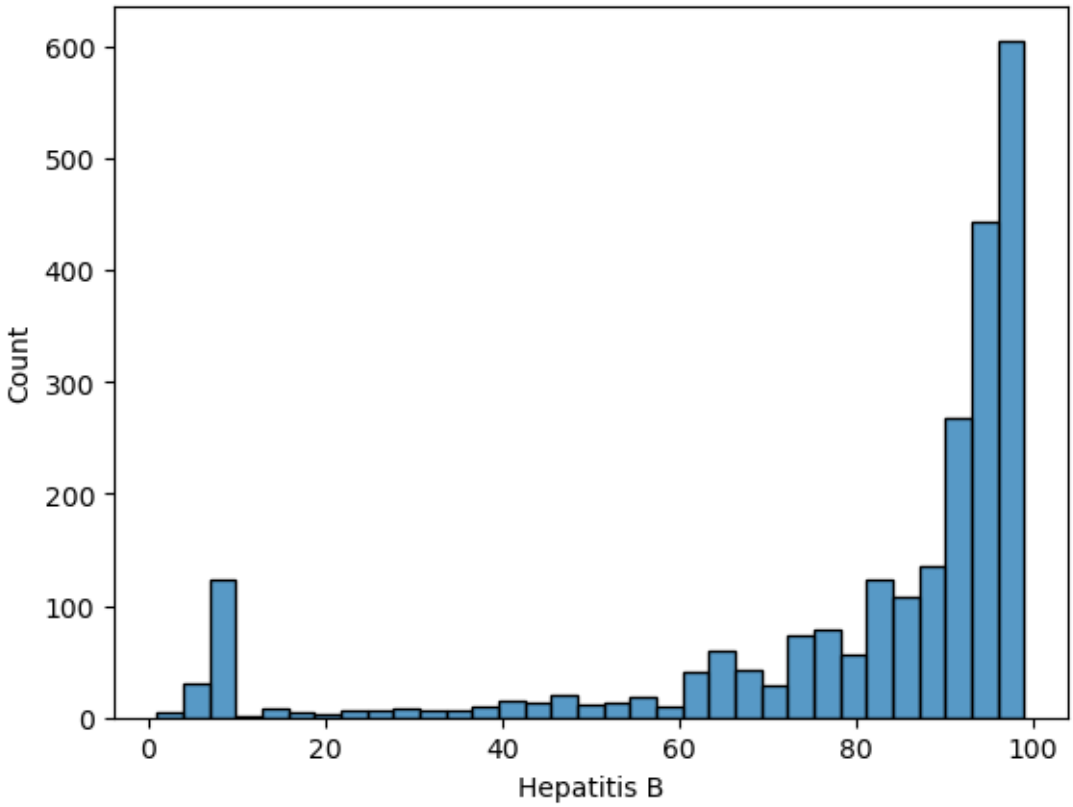

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



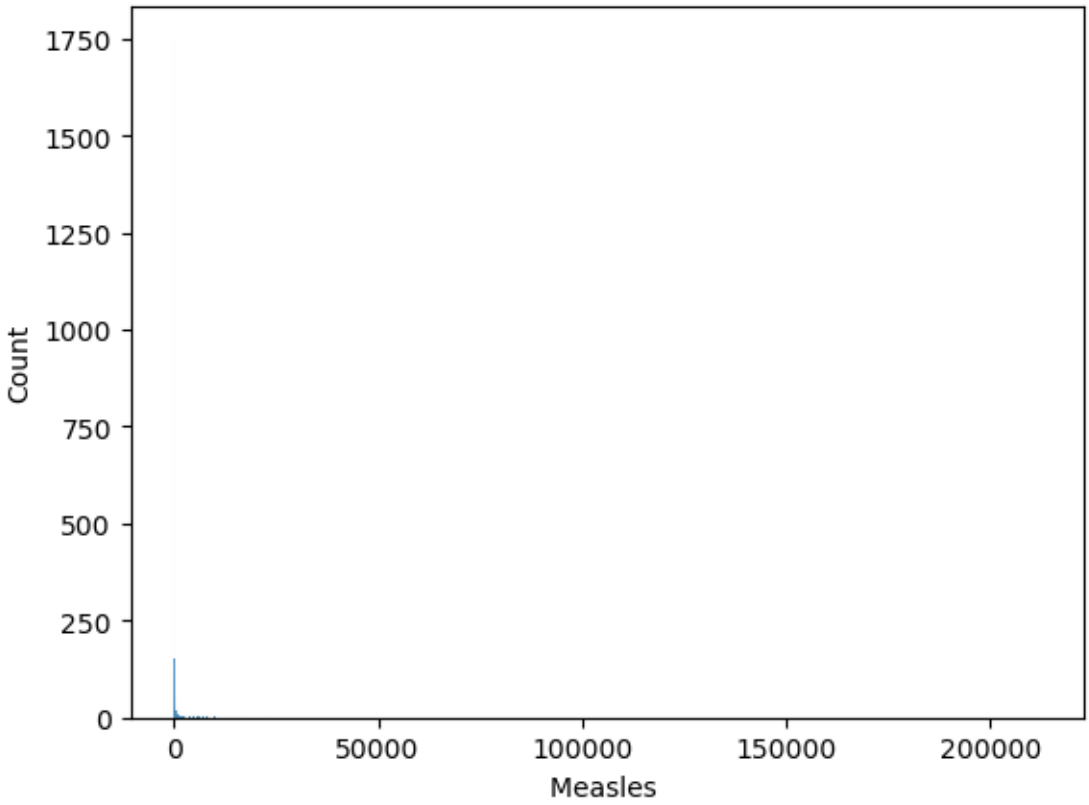

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



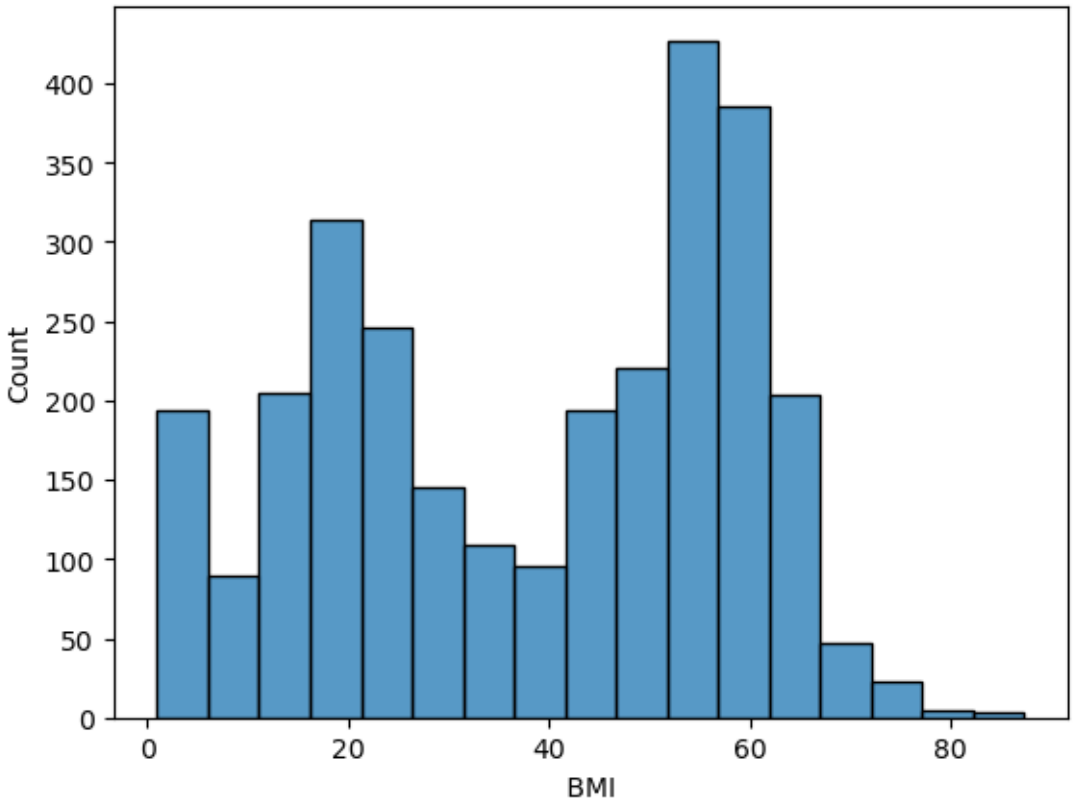

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



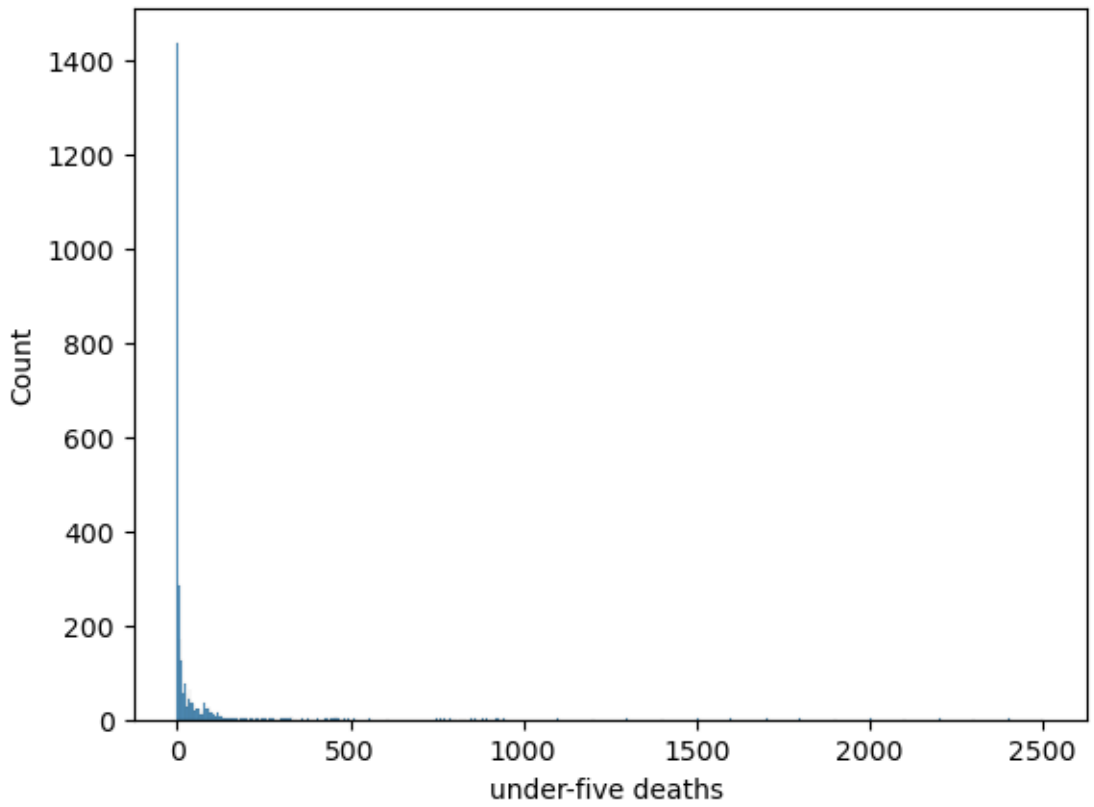

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



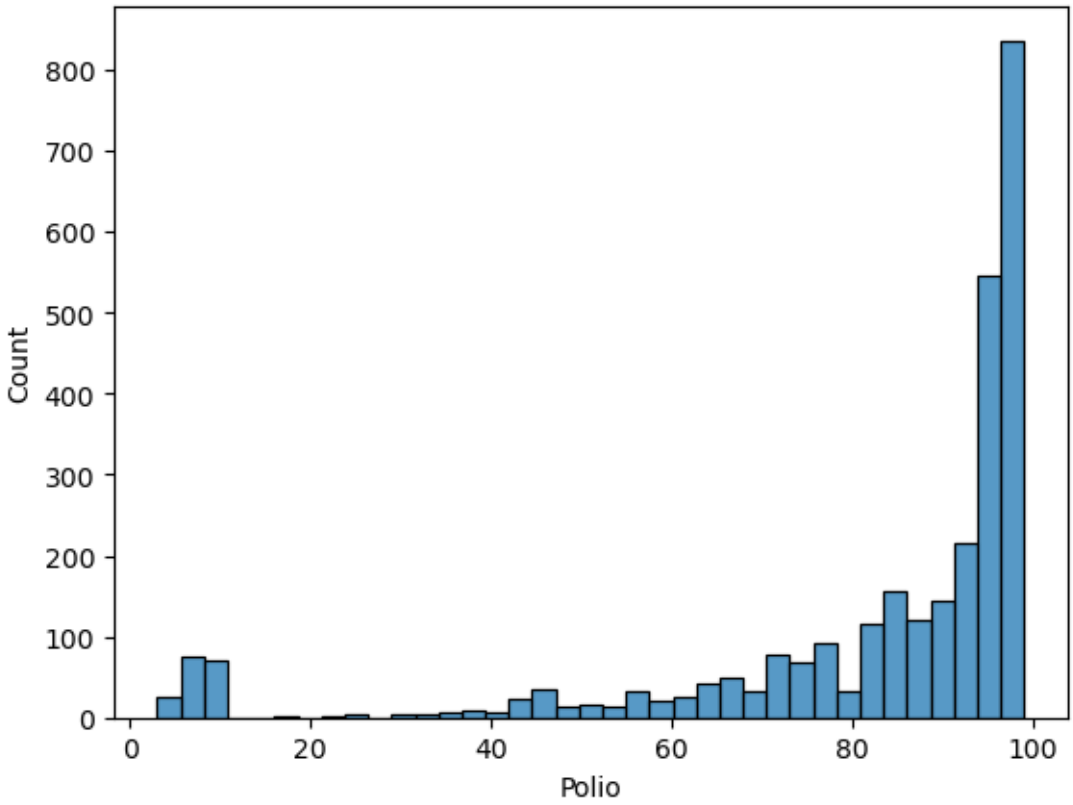

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



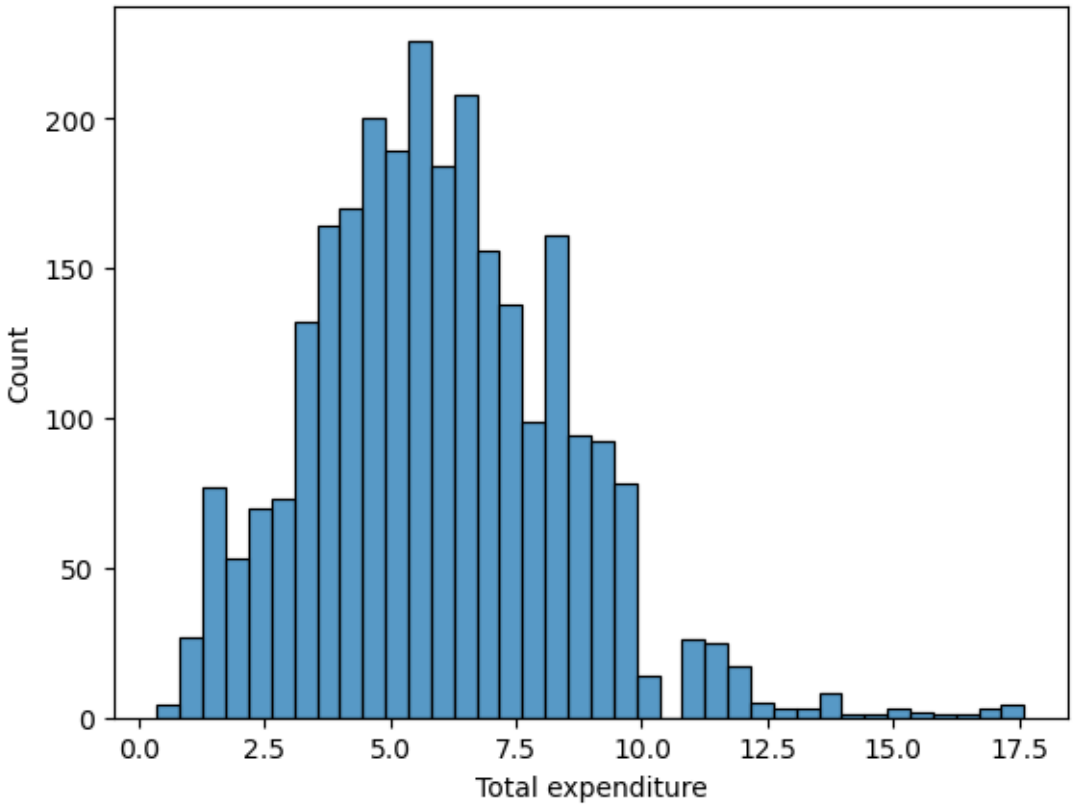

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



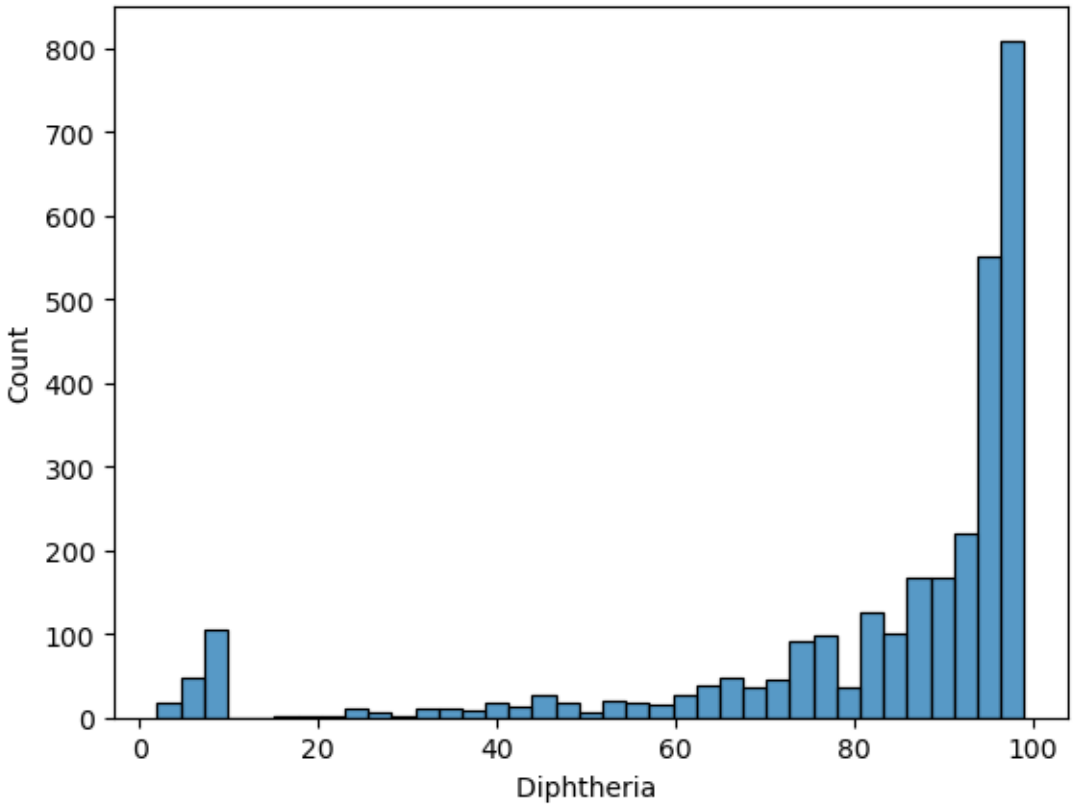

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



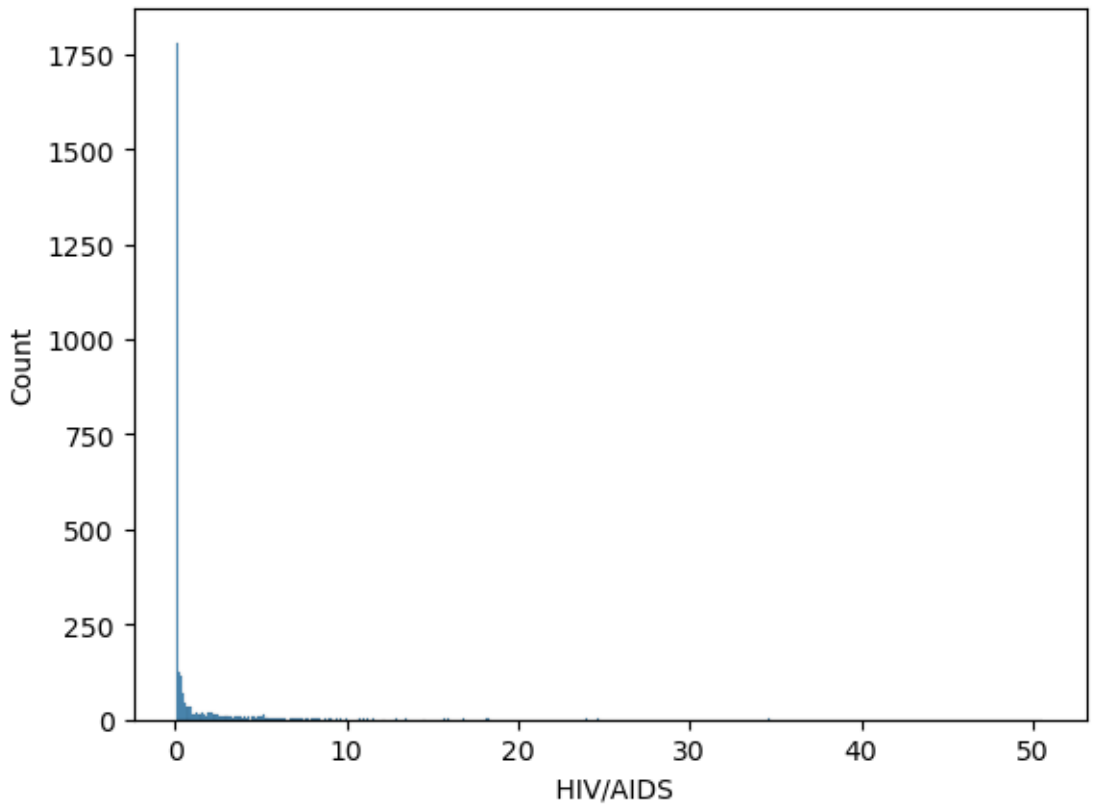

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



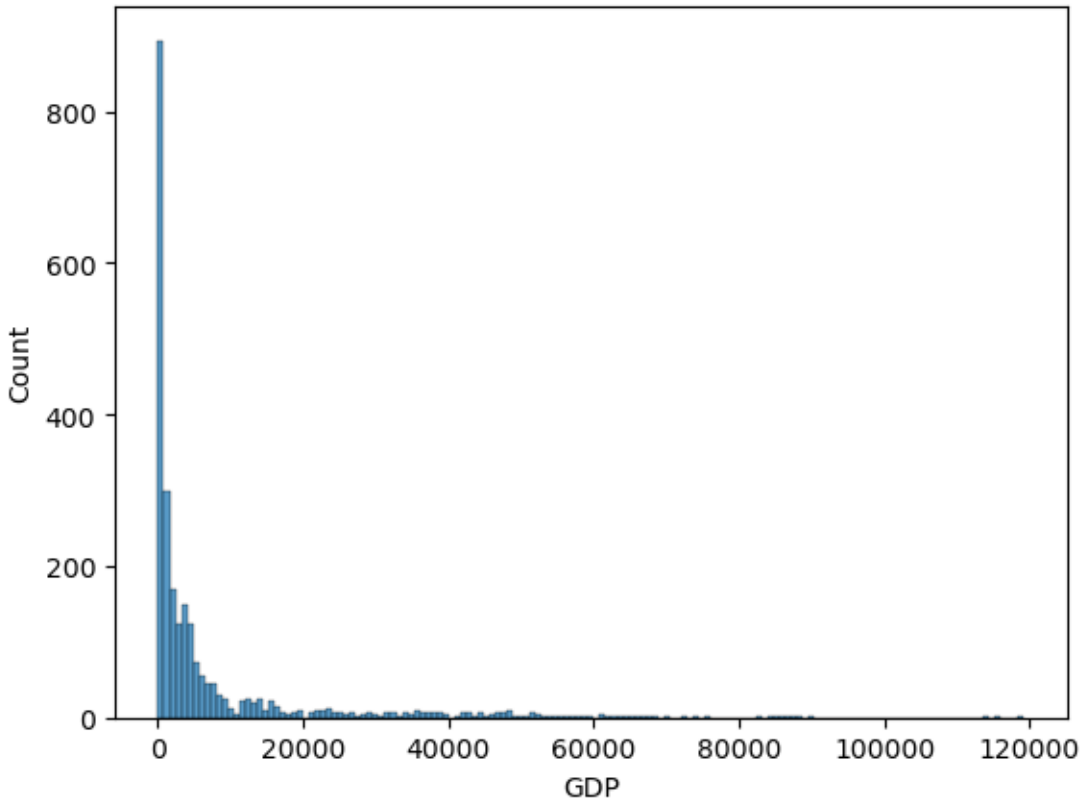

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



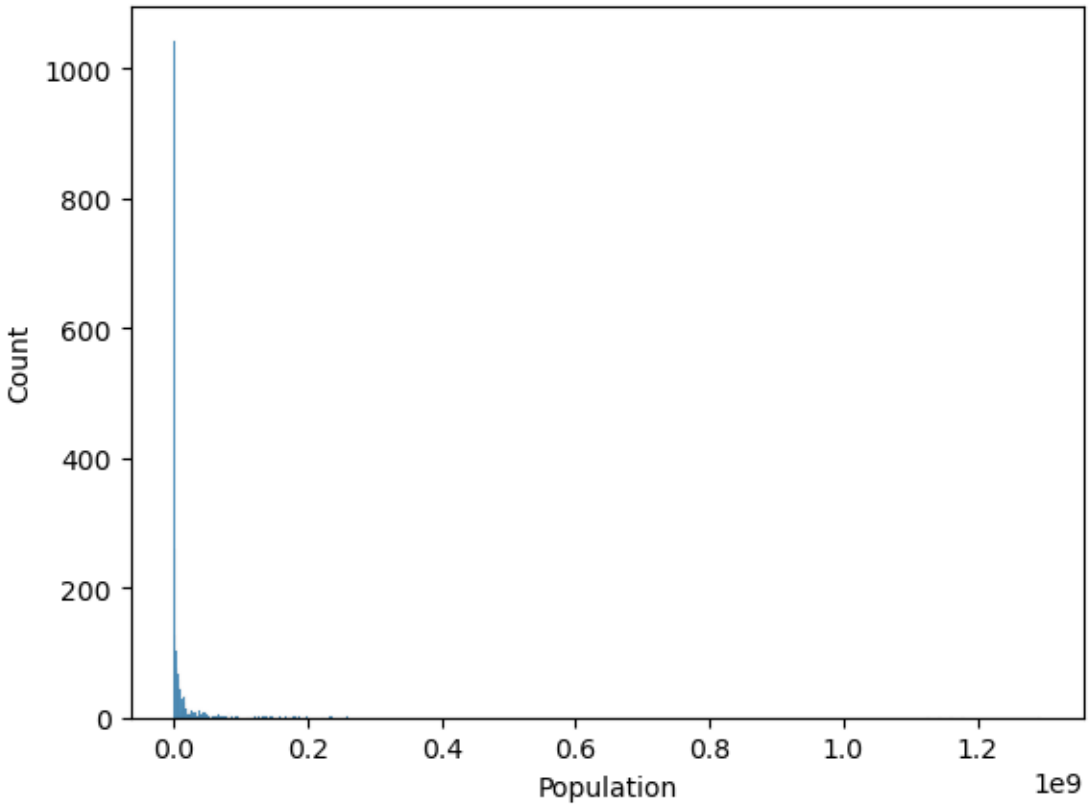

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



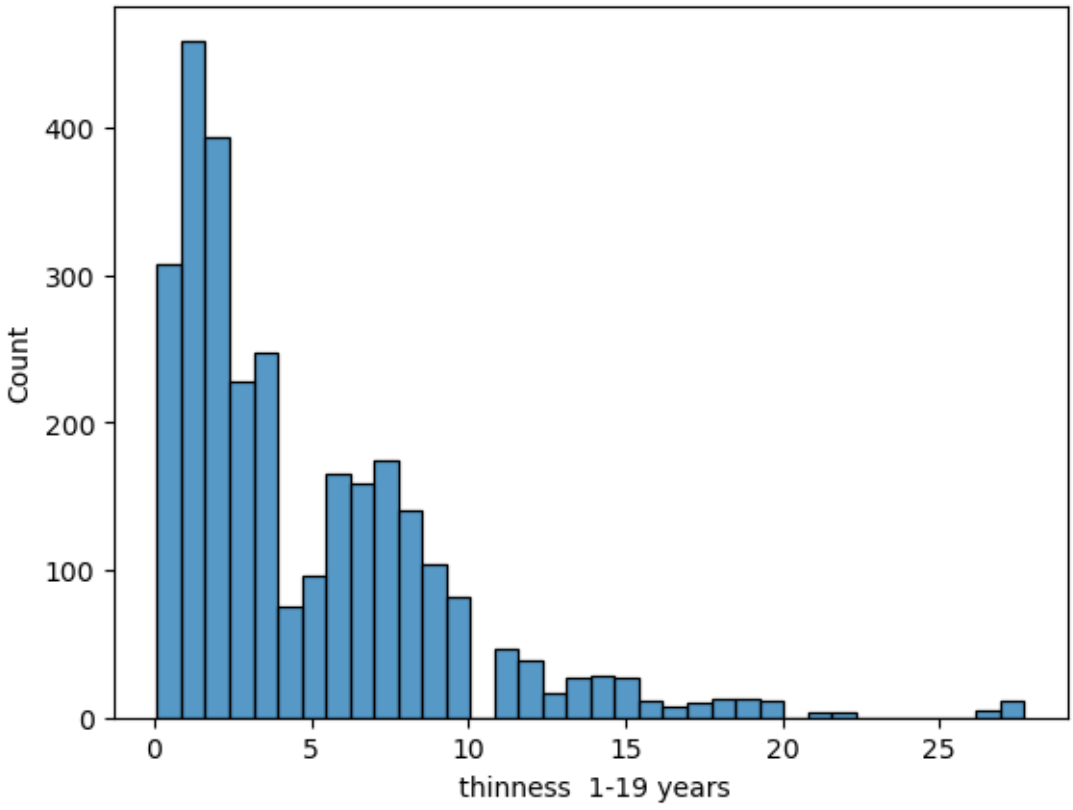

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



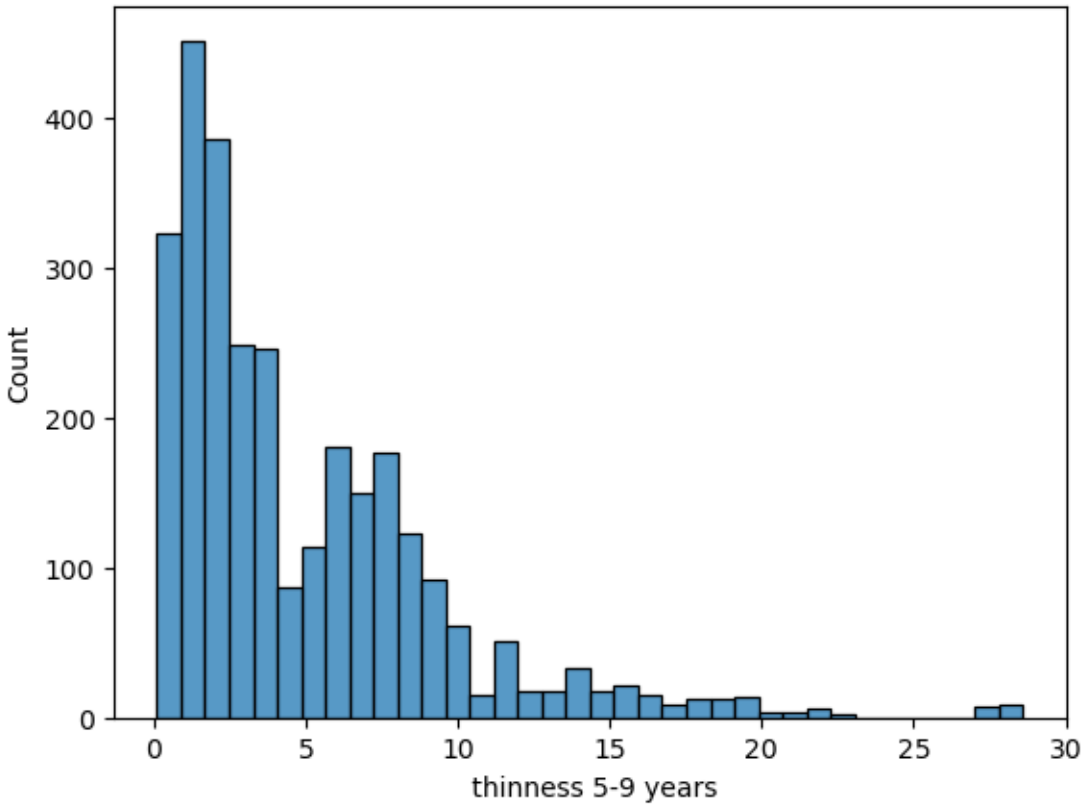

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



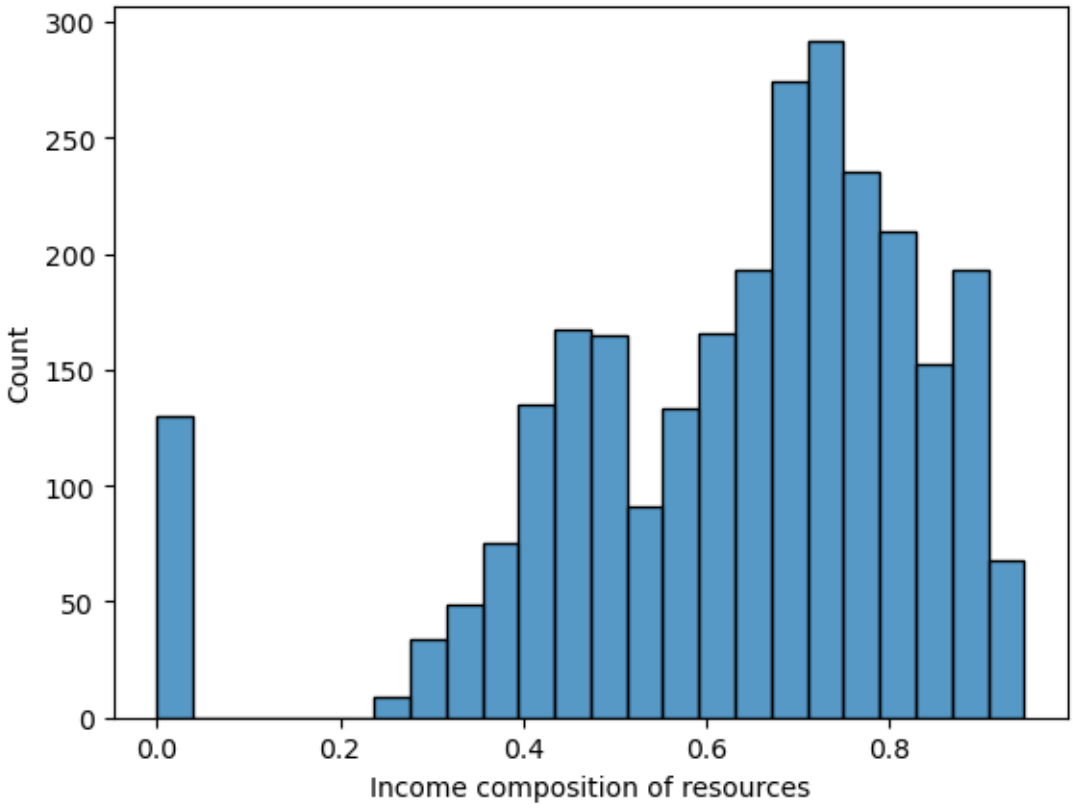

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



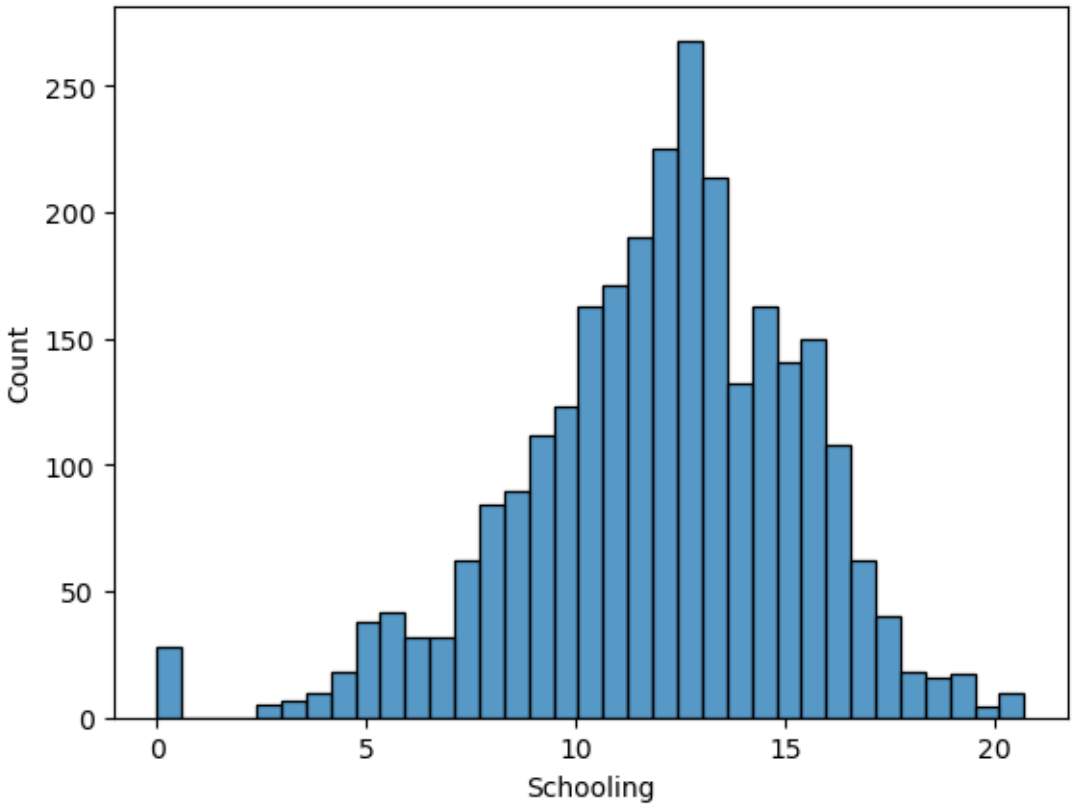

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



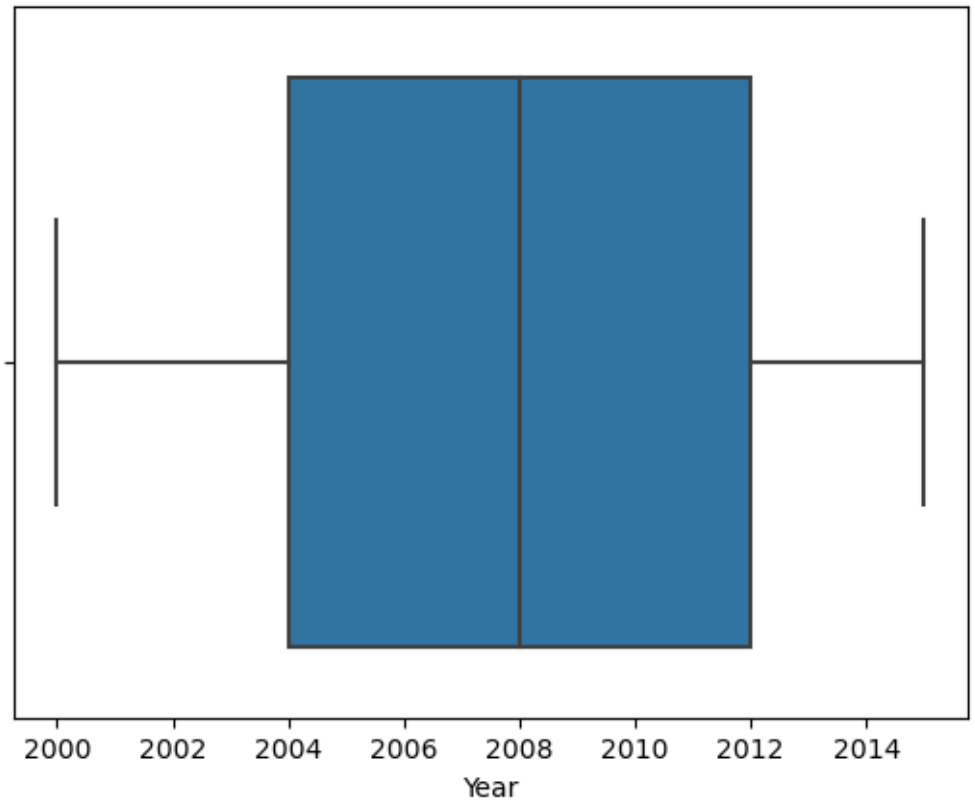

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

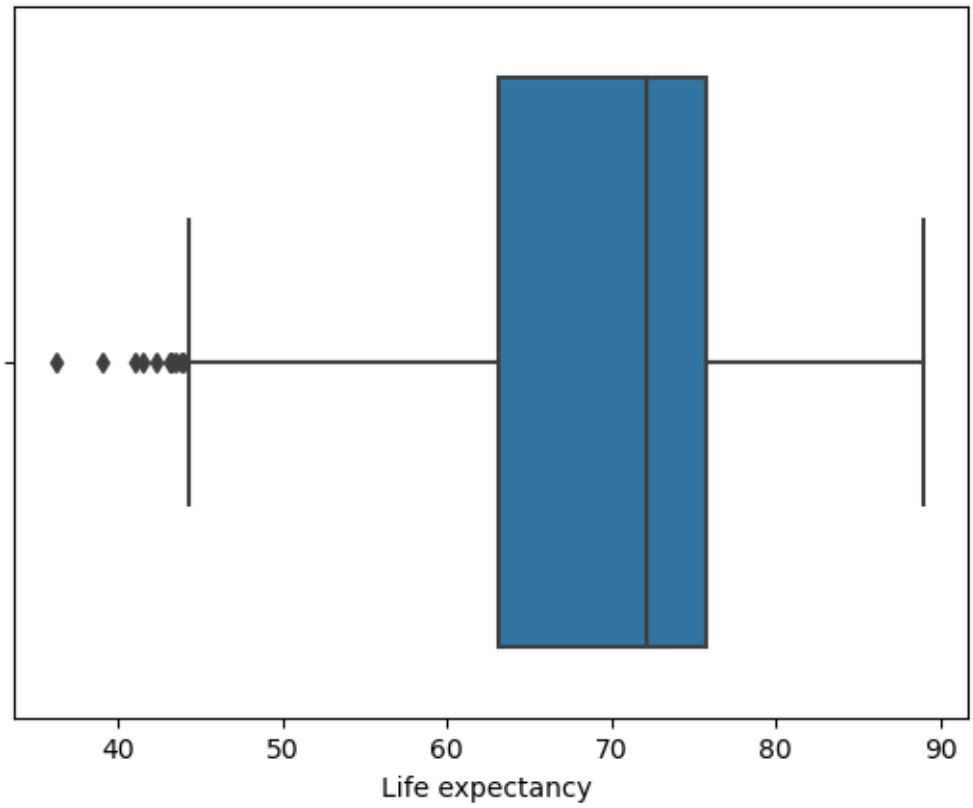


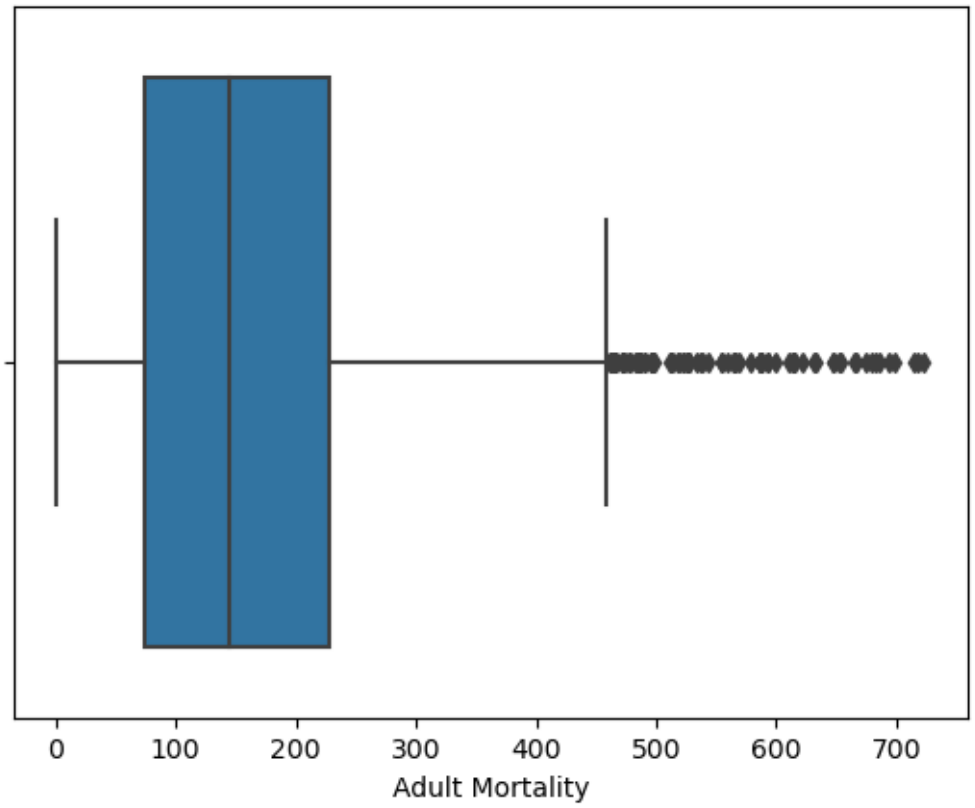

```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

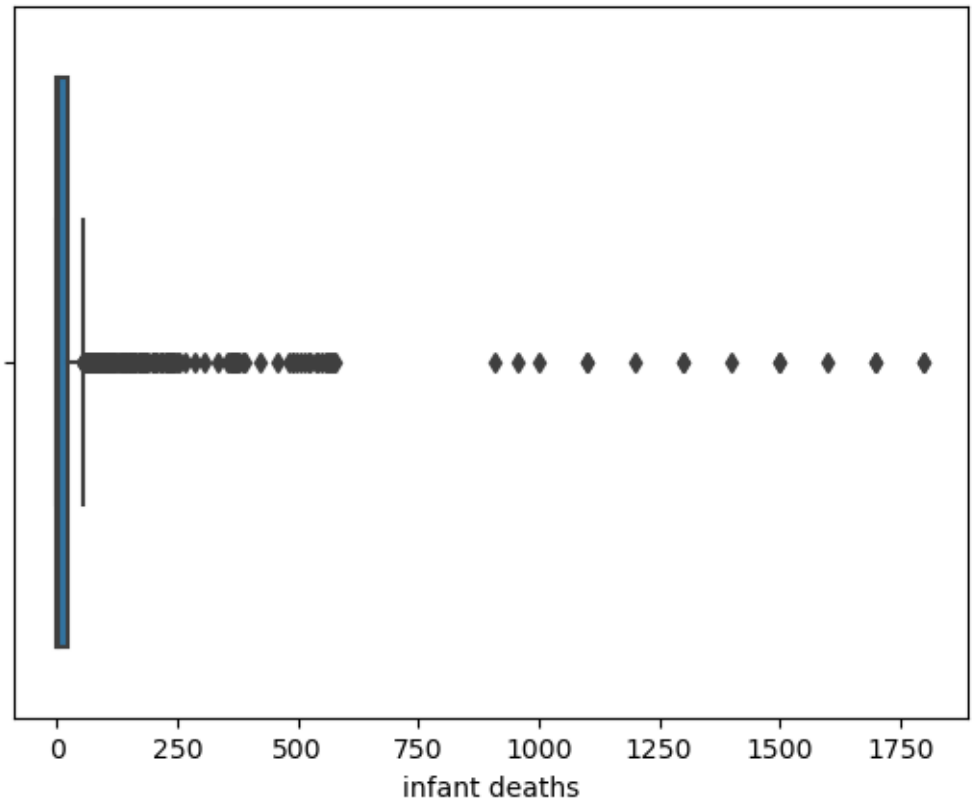


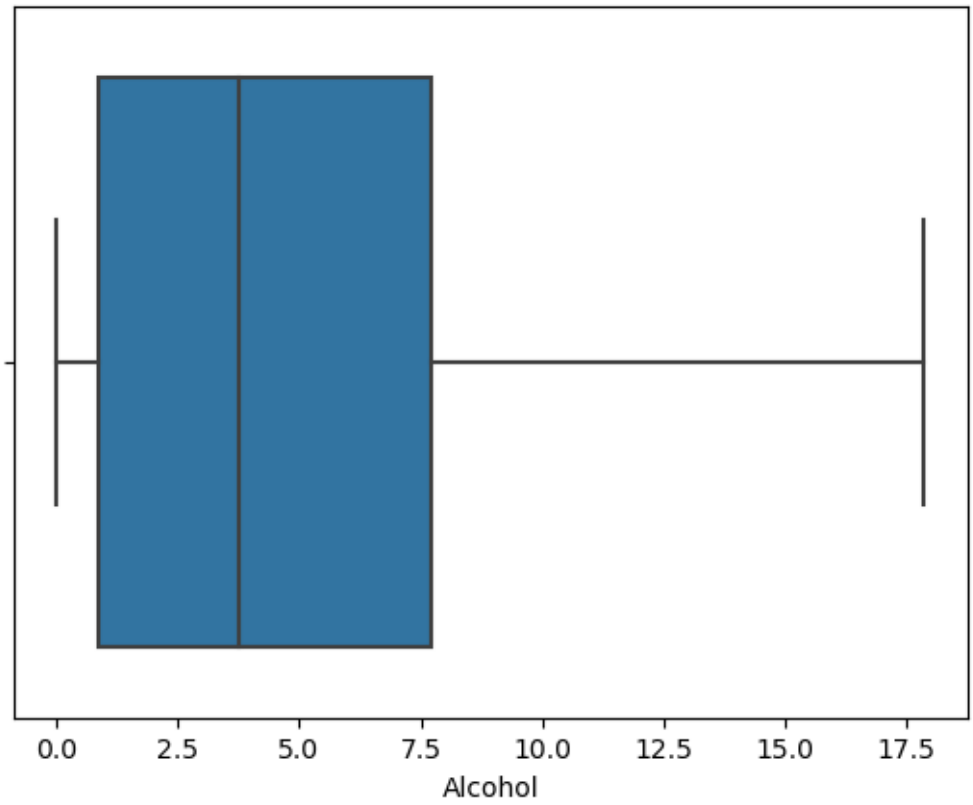
[43]:
```
## Boxplot to identify the outlier
for i in data.select_dtypes(include="number").columns:
    sns.boxplot(data=data, x=i)
    plt.show()
```

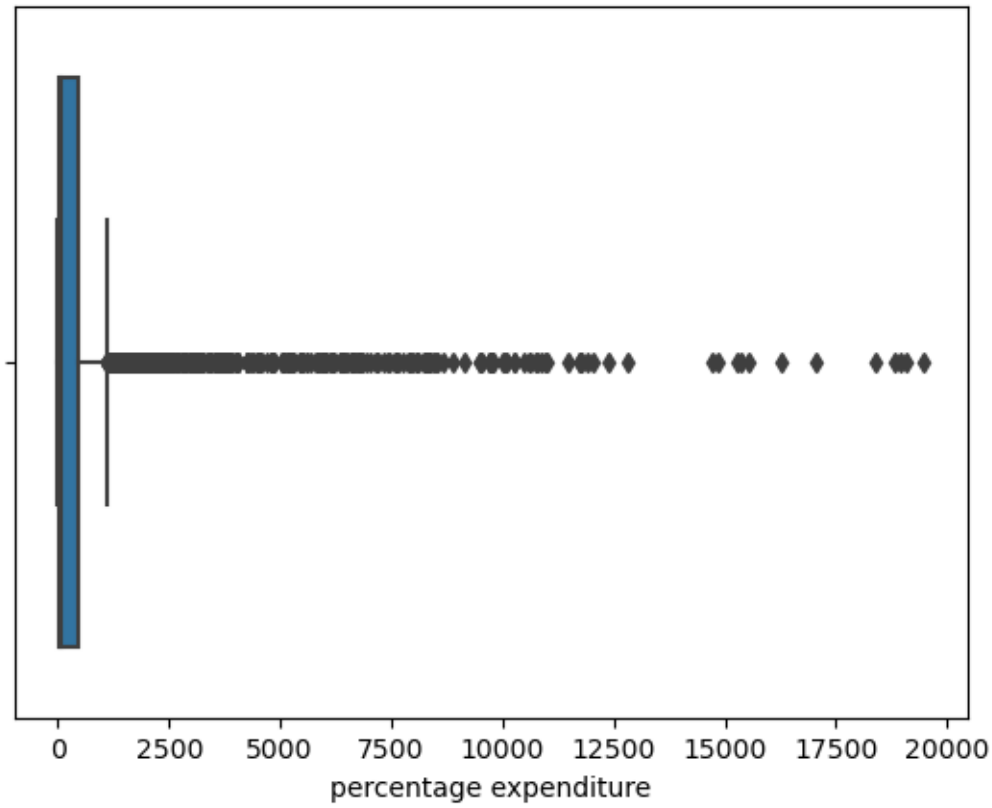

2000
2002
2004
2006
2008
2010
2012
2014
Year

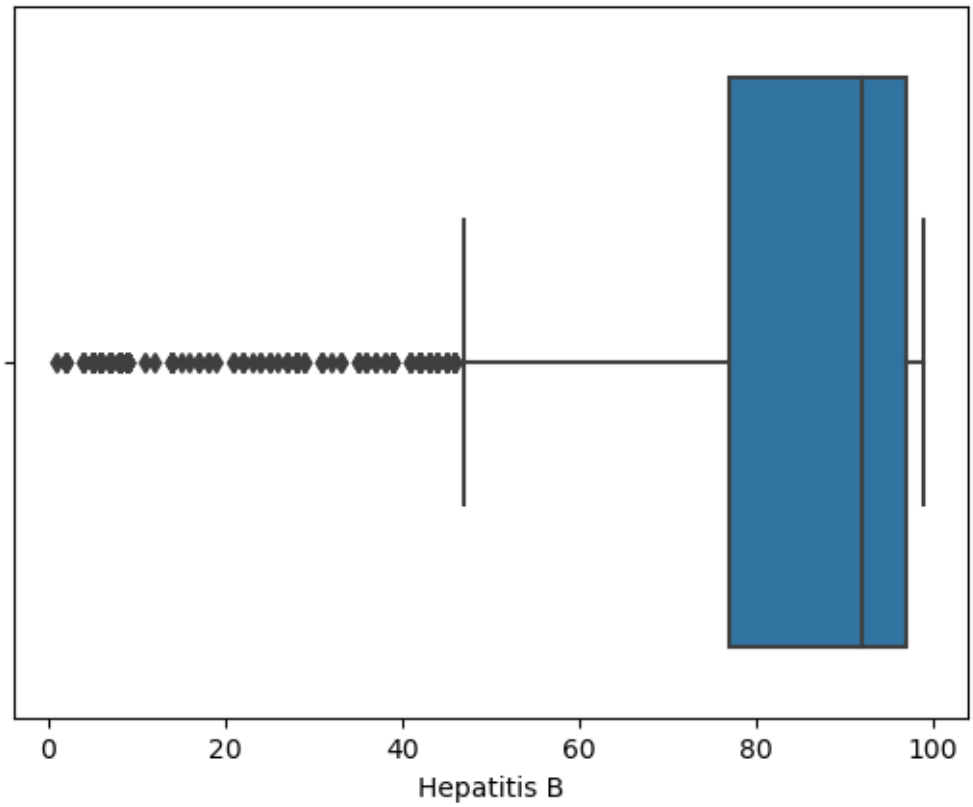

40
50
60
70
80
90
Life expectancy

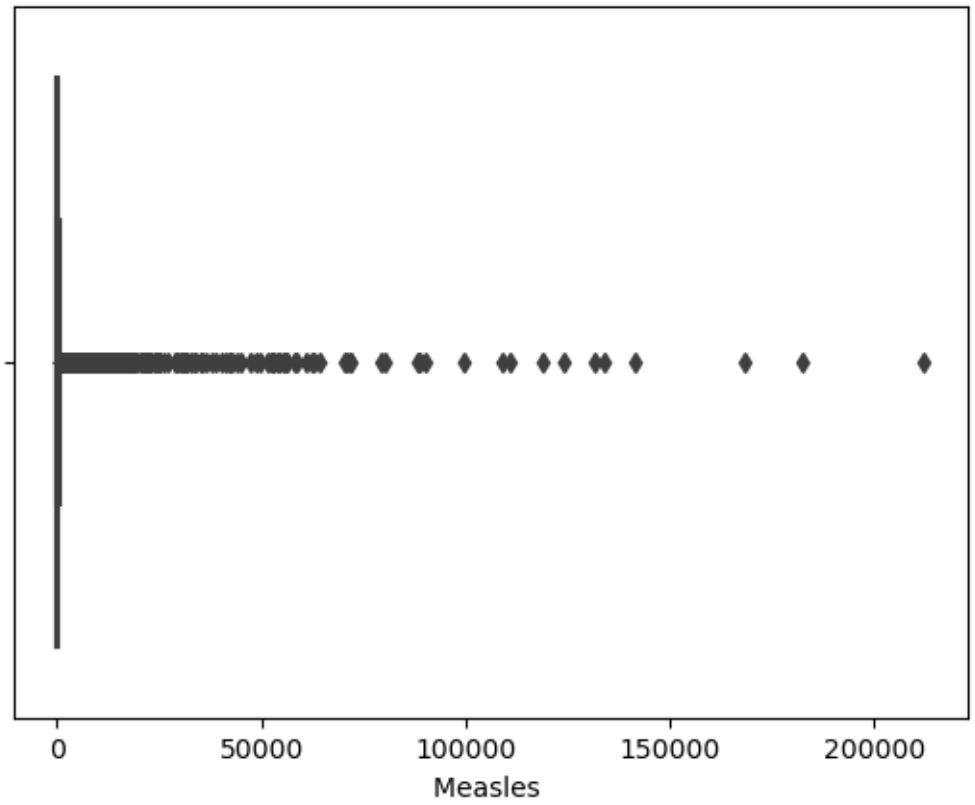

0
100
200
300
400
500
600
700
Adult Mortality

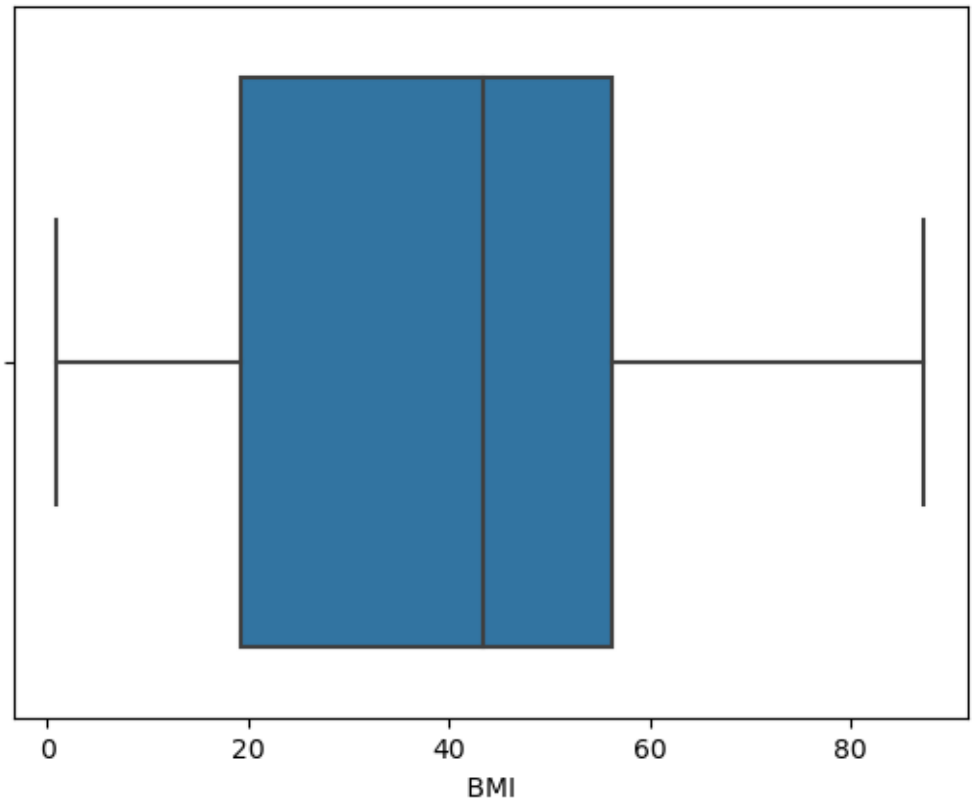

0
250
500
750
1000
1250
1500
1750
infant deaths

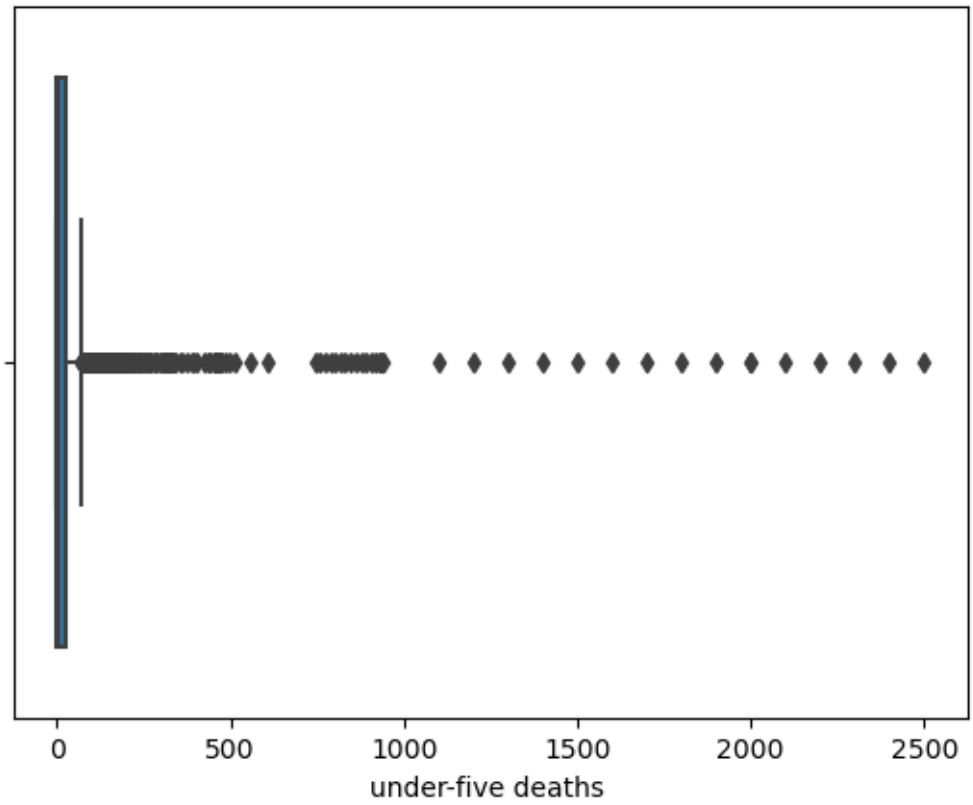

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol

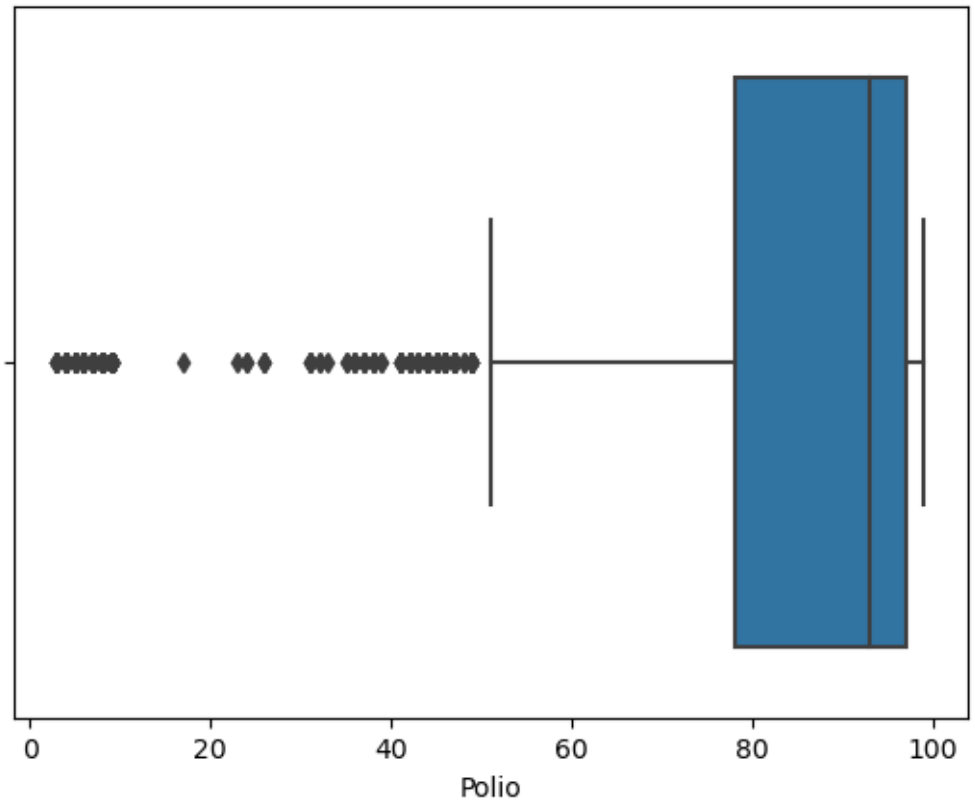

0
2500
5000
7500
10000
12500
15000
17500
20000
percentage expenditure

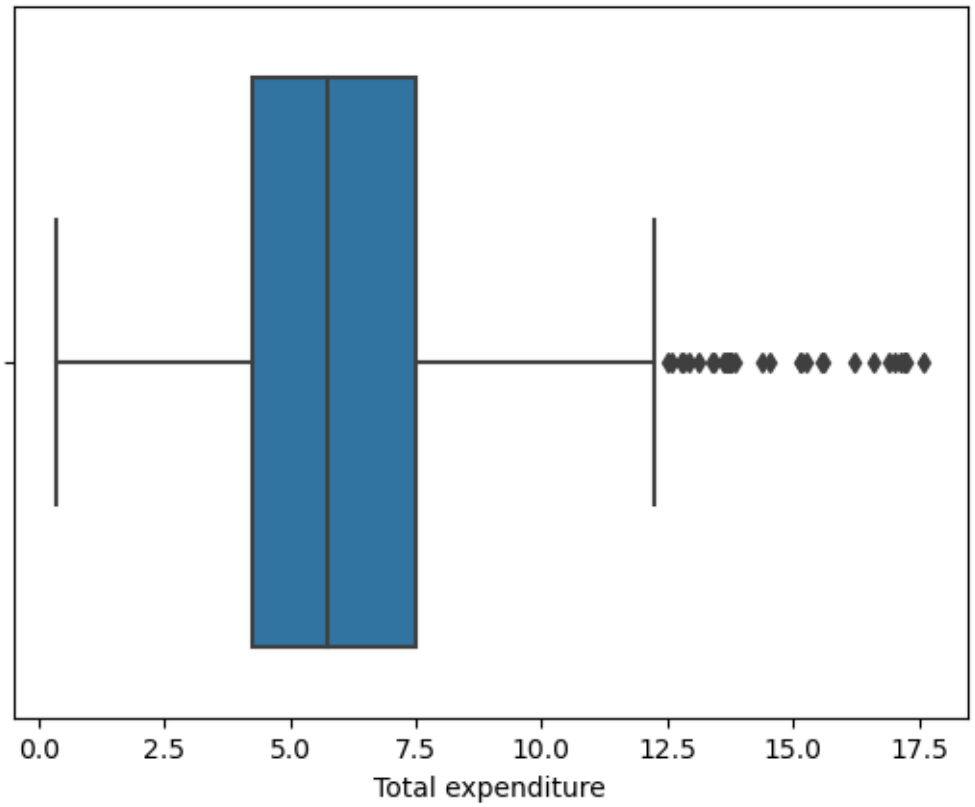

0
20
40
60
80
100
Hepatitis B

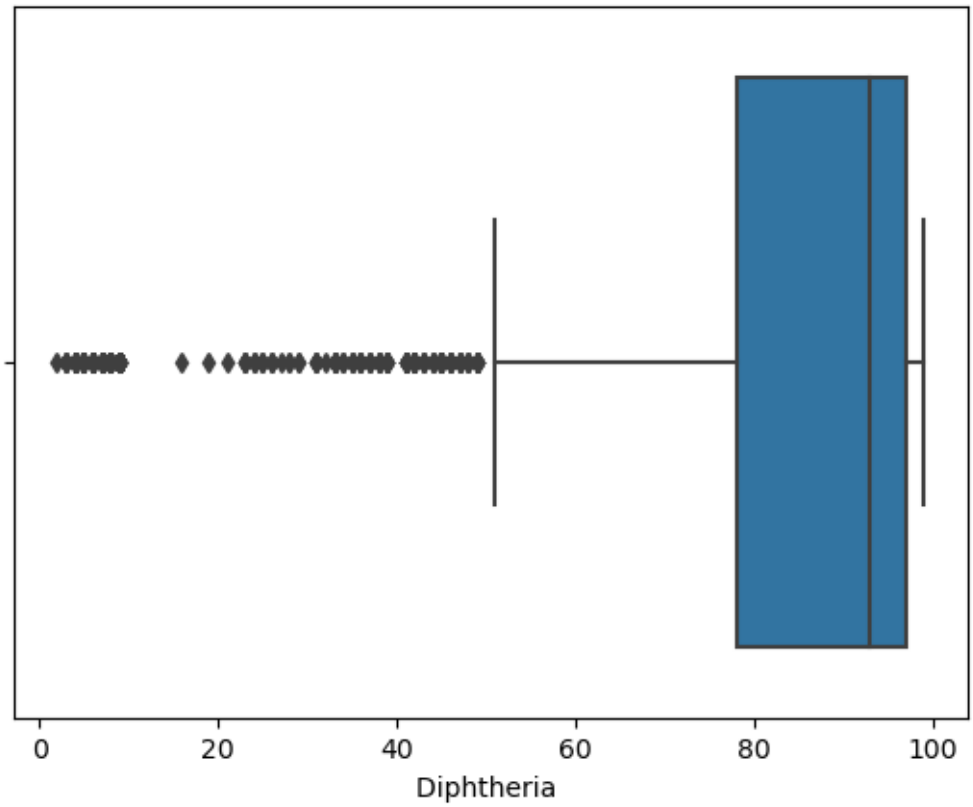

0
50000
100000
150000
200000
Measles

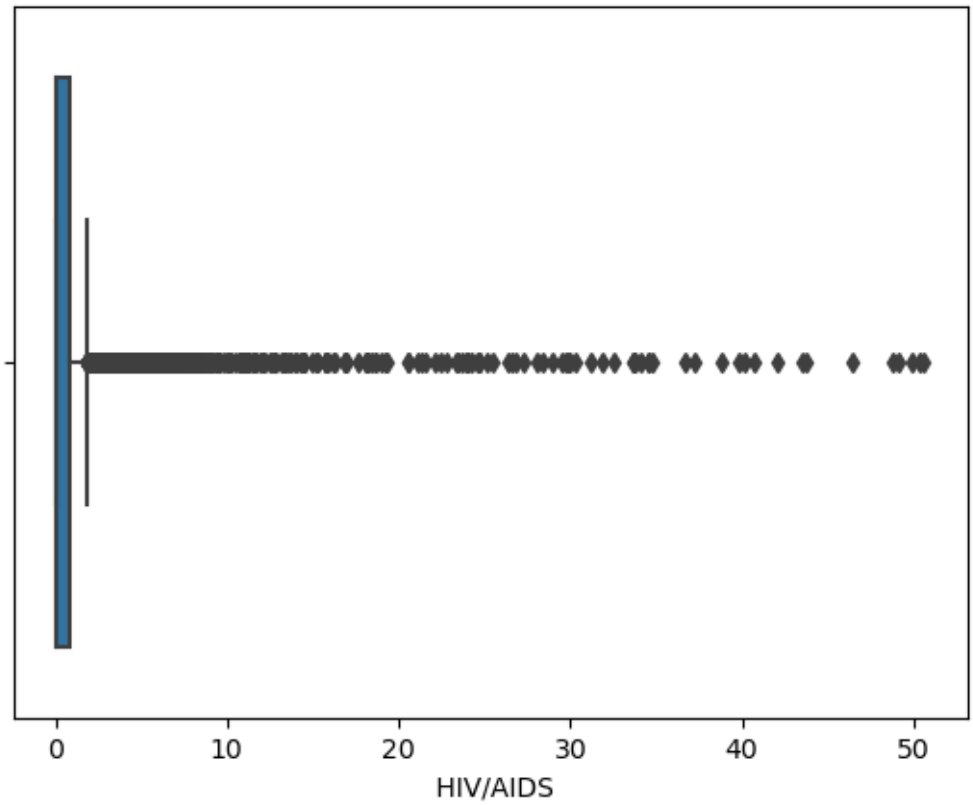

0
20
40
60
80
BMI

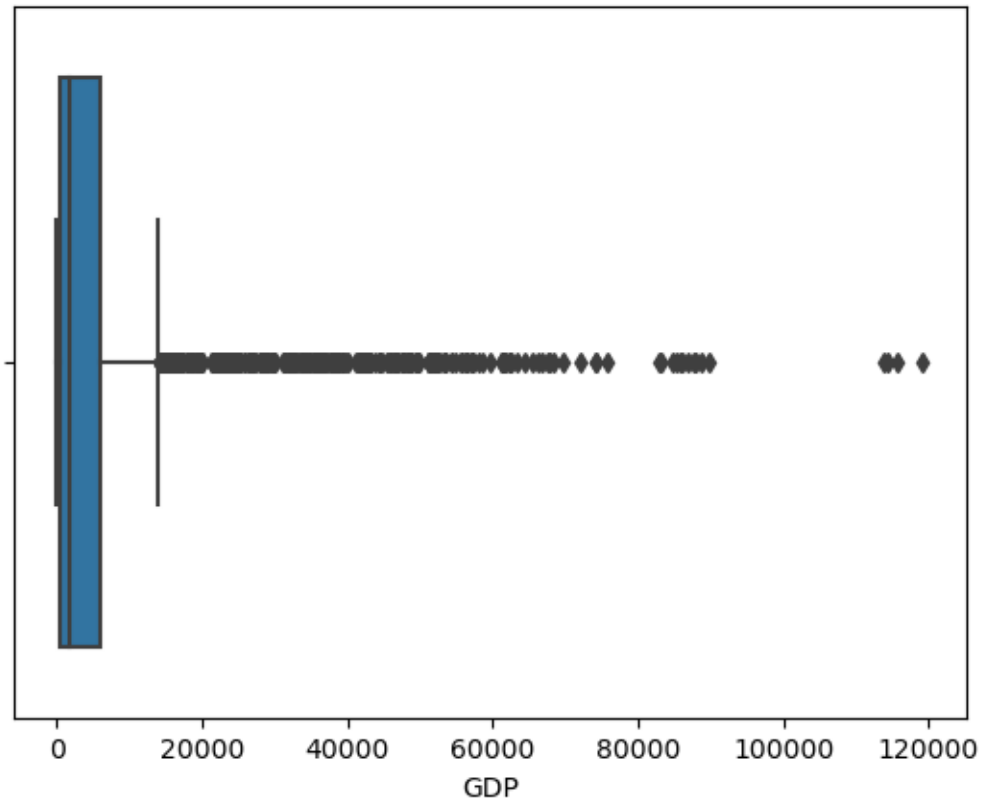

0
500
1000
1500
2000
2500
under-five deaths

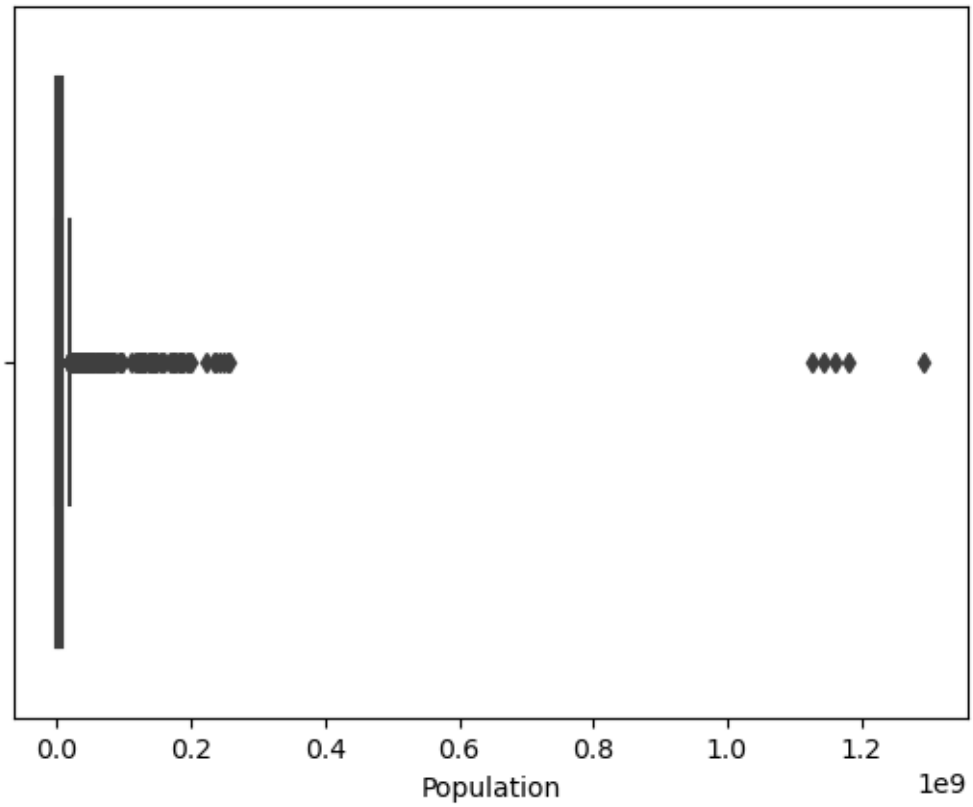

0
20
40
60
80
100
Polio

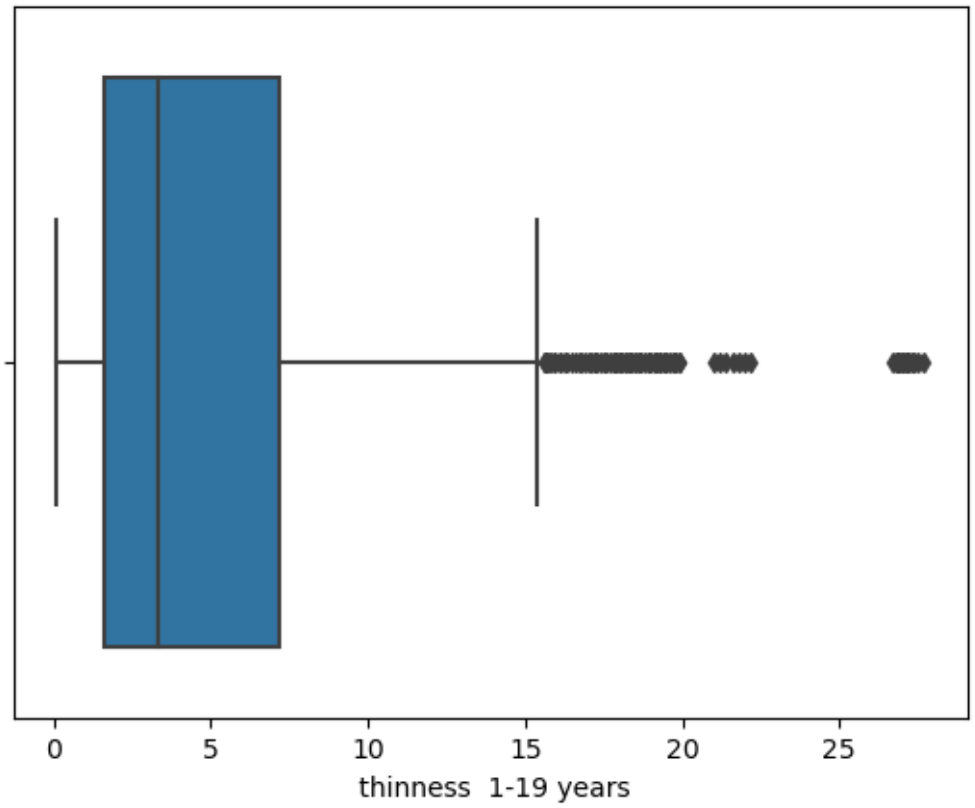

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Total expenditure

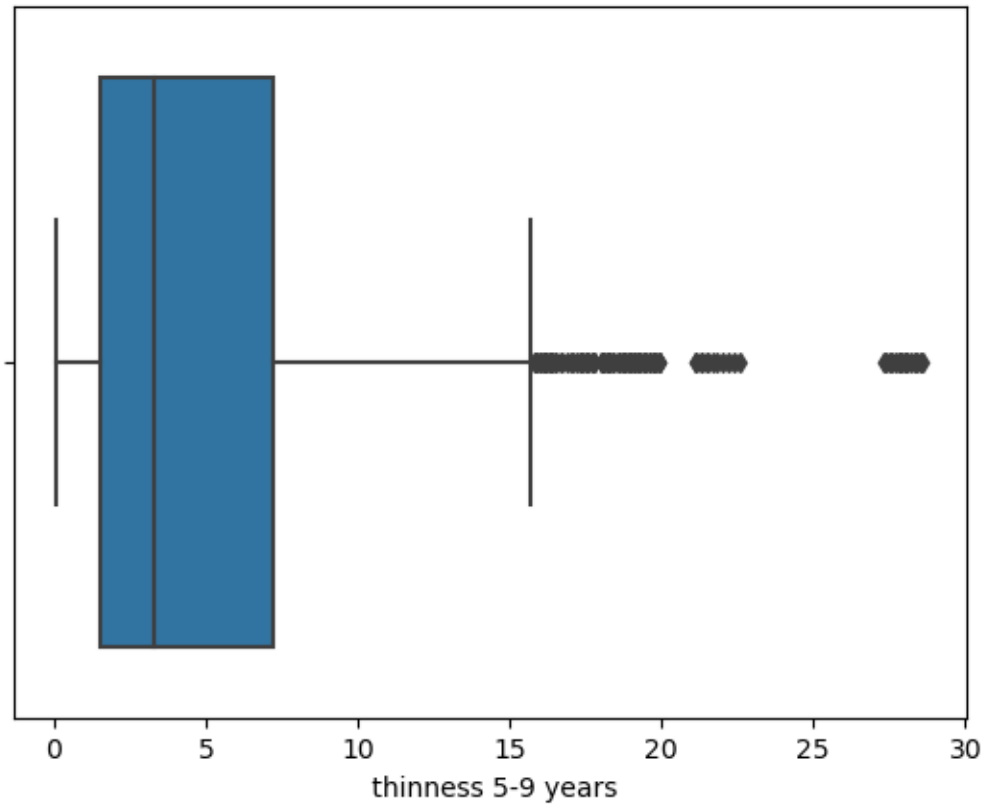

0
20
40
60
80
100
Diphtheria

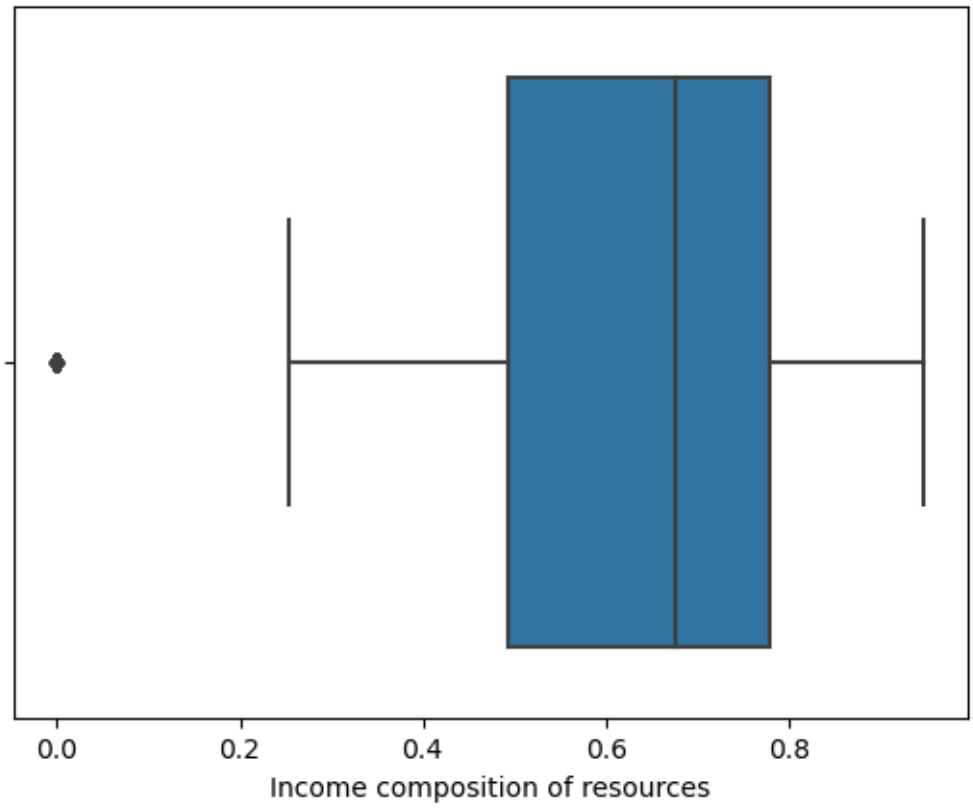

0
10
20
30
40
50
HIV/AIDS

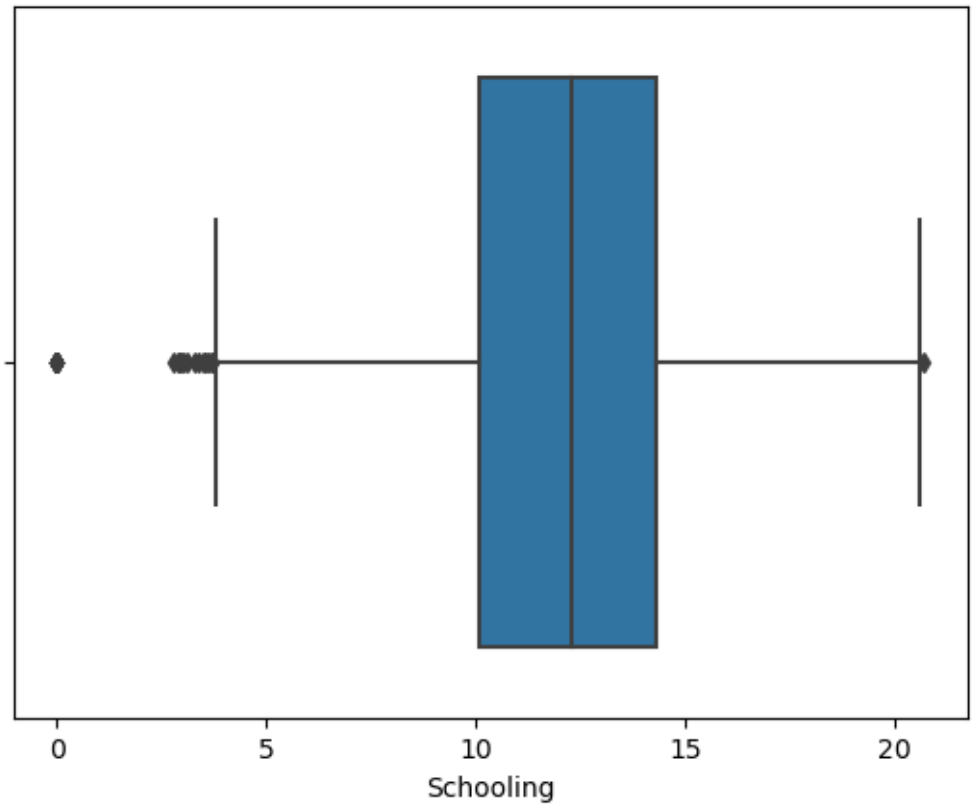

0
20000
40000
60000
80000
100000
120000
GDP


0.0
0.2
0.4
0.6
0.8
1.0
1.2
Population
1e9


0
5
10
15
20
25
thinness 1-19 years


0
5
10
15
20
25
30
thinness 5-9 years


0.0
0.2
0.4
0.6
0.8
Income composition of resources

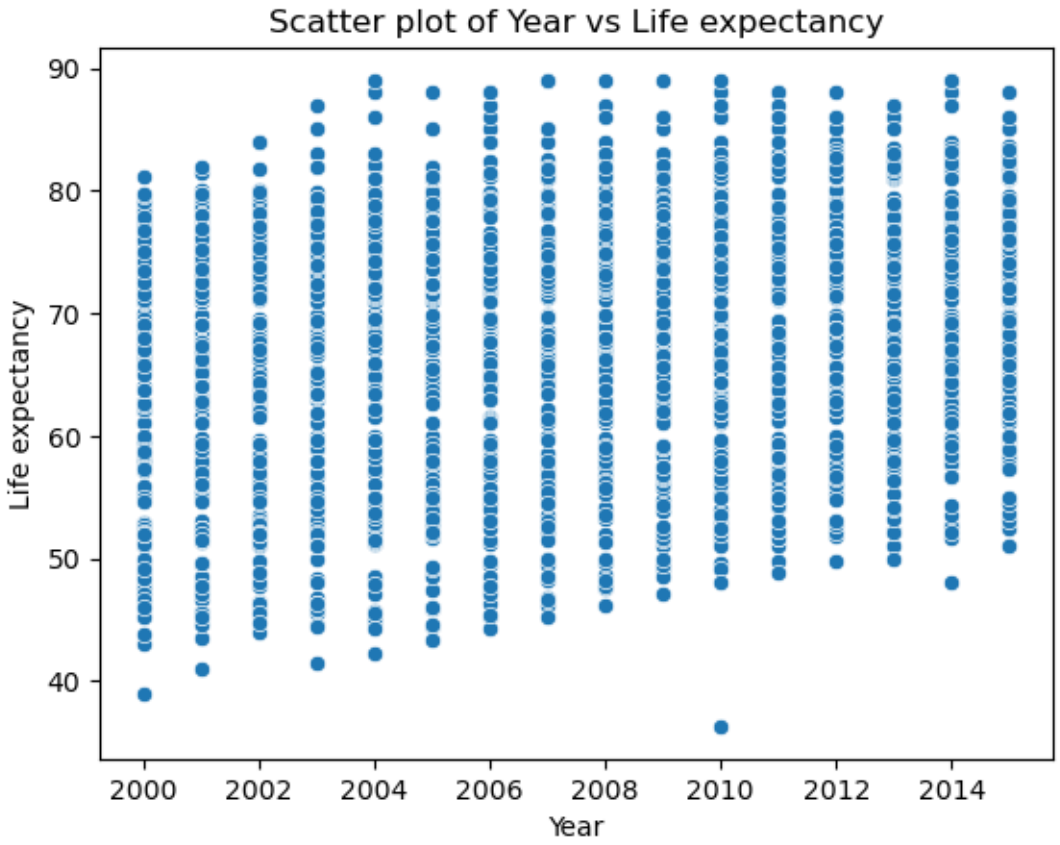


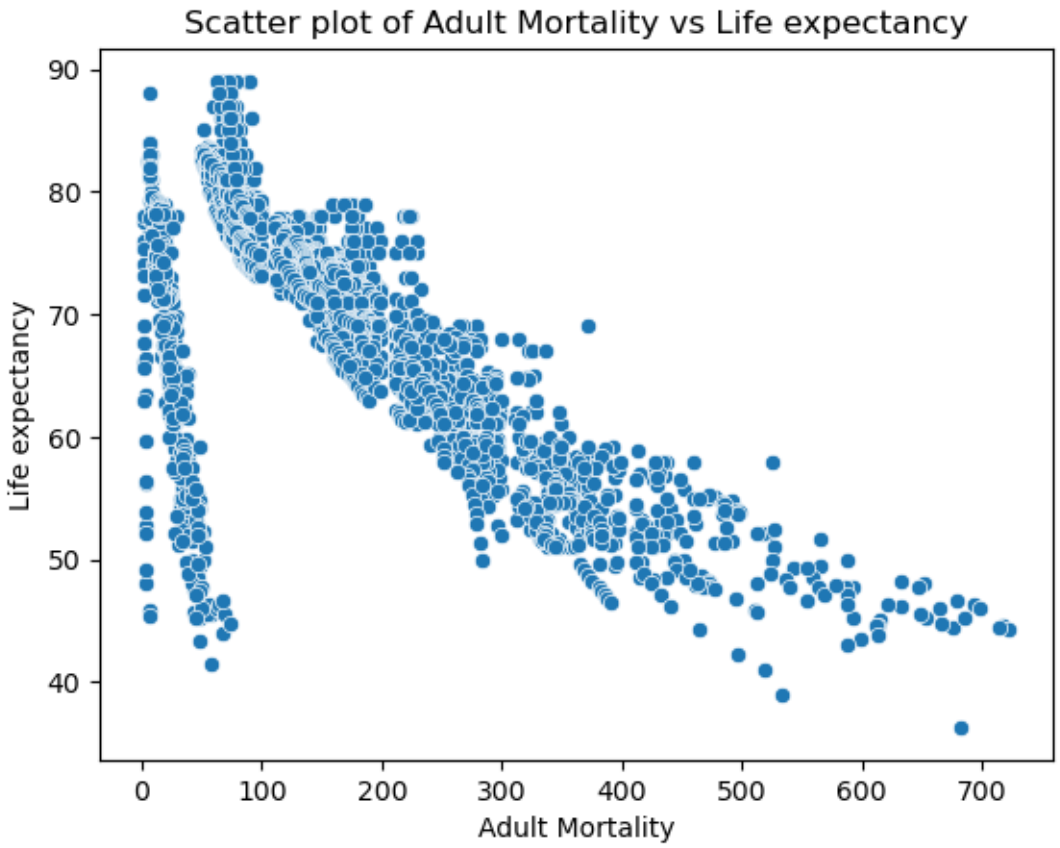
[47]:
```
data.rename(columns={'Measles ': 'Measles'}, inplace=True)
```

[50]:
```
print(data.columns)
data.rename(columns={' thinness  1-19 years': 'thinness_1_19_years'},␣
 ↪inplace=True)
data.columns = data.columns.str.replace('\s+', ' ', regex=True).str.strip()
data.columns = data.columns.str.replace('\s+', ' ', regex=True).str.strip()
print(data.columns)  # Check cleaned column names
```

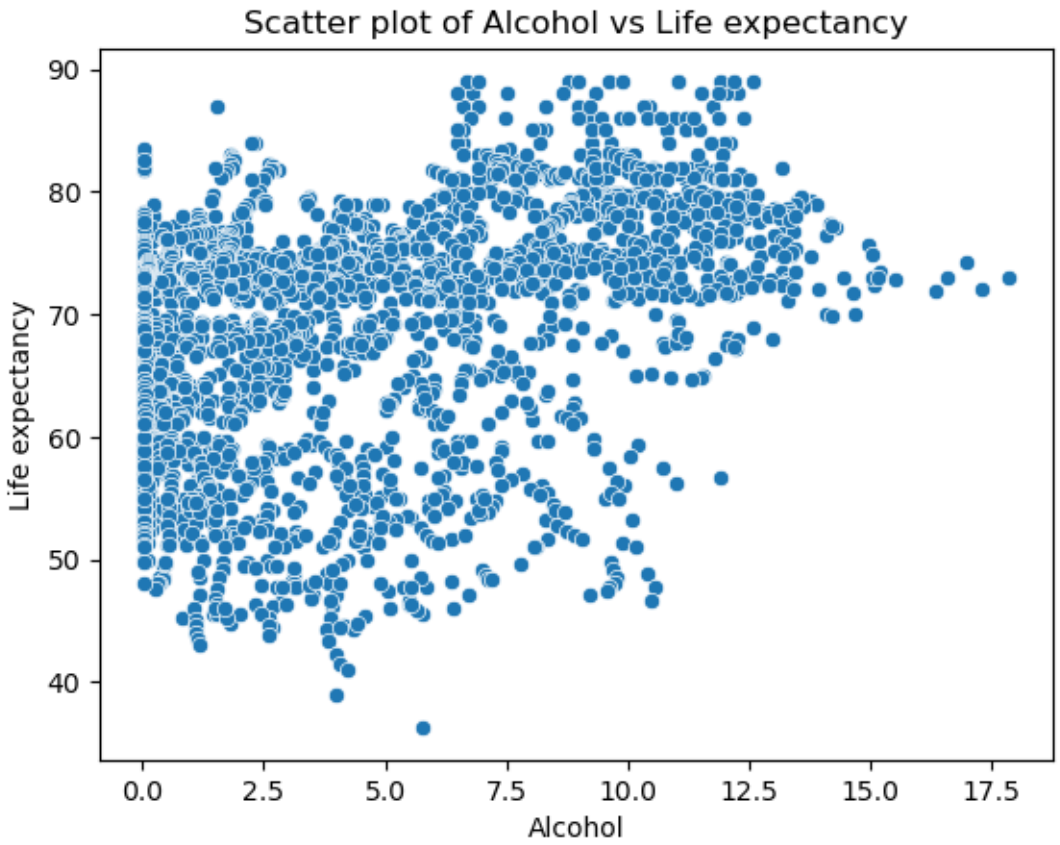
```
Index(['Country', 'Year', 'Status', 'Life expectancy', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B',
       'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total expenditure',
       'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness  1-19 years',
       'thinness 5-9 years', 'Income composition of resources', 'Schooling'],
      dtype='object')
Index(['Country', 'Year', 'Status', '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')
```

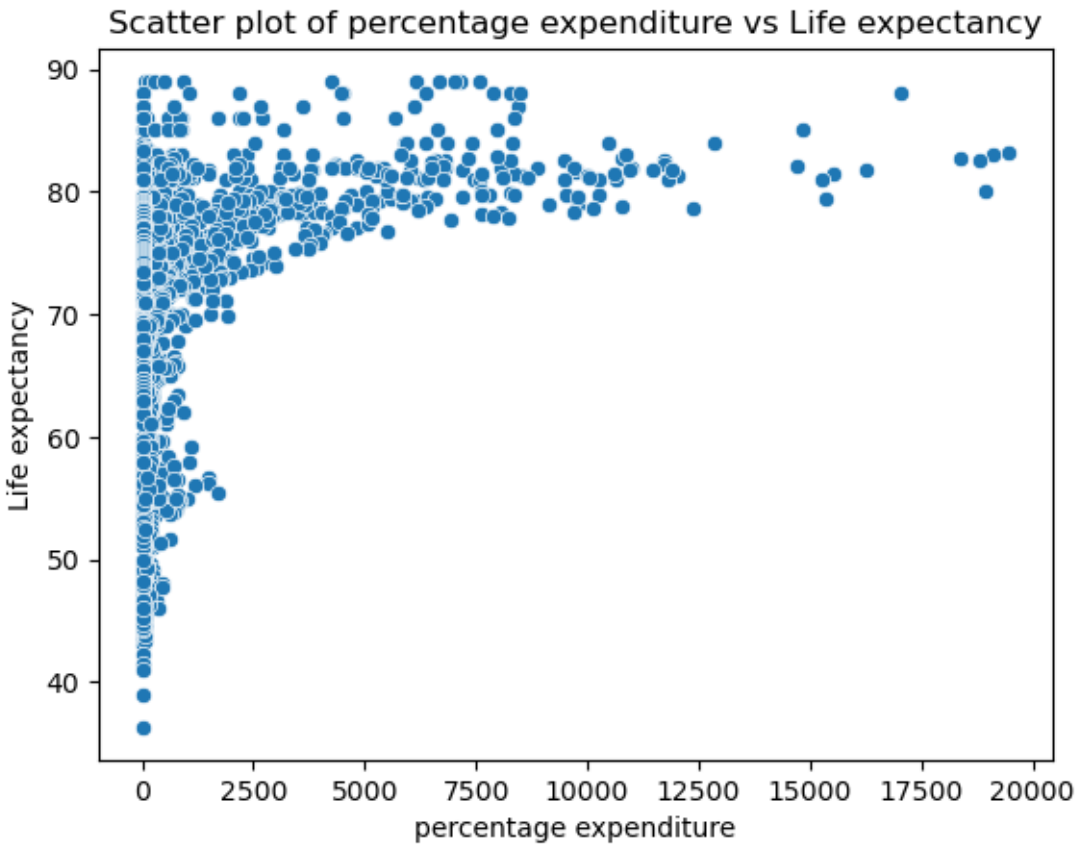
[51]:
```
## scatterplot to understand the relationship
# Strip any leading/trailing whitespace from column names
data.columns = data.columns.str.strip()

# List of columns to create scatter plots for
columns = ['Year', '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']

# Loop through each column and create a scatter plot against 'Life expectancy'
for i in columns:
    sns.scatterplot(data=data, x=i, y='Life expectancy')  # Make sure 'Life␣
 ↪expectancy' is correct
    plt.title(f'Scatter plot of {i} vs Life expectancy')
    plt.show()
```

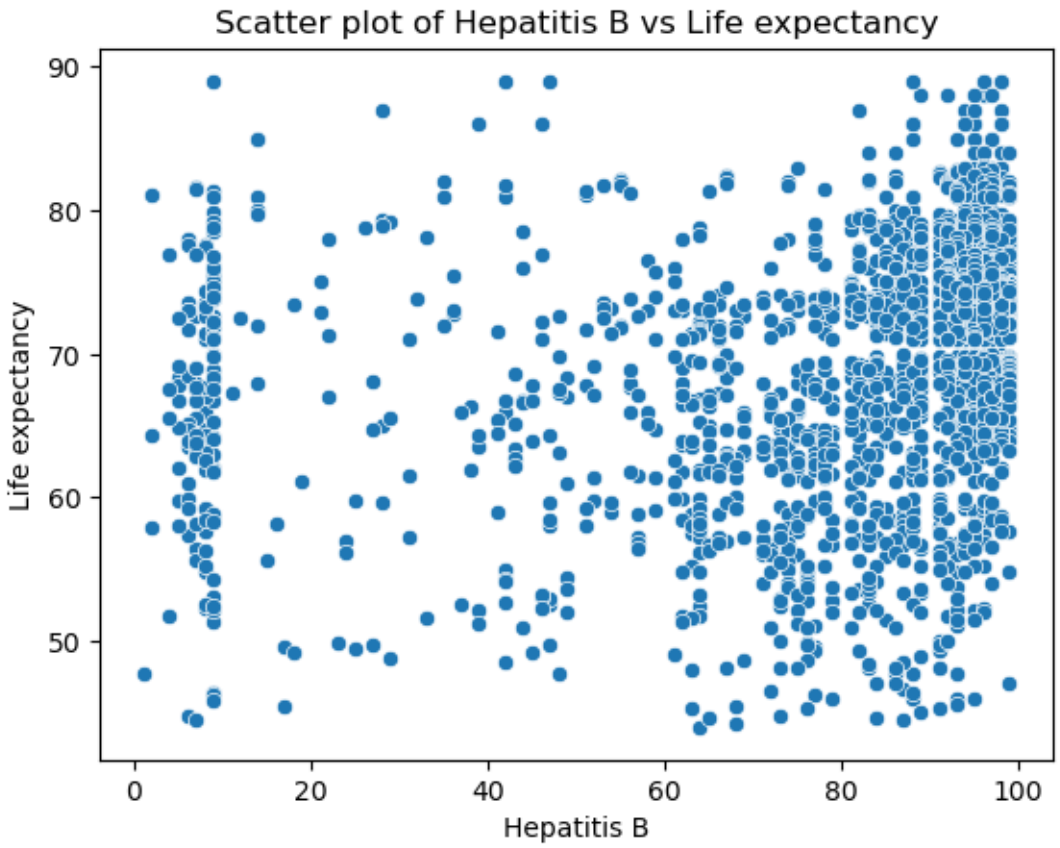



Scatter plot of Adult Mortality vs Life expectancy
Life expectancy
Adult Mortality
90
80
70
60
50
40
0
100
200
300
400
500
600
700


Scatter plot of infant deaths vs Life expectancy
Life expectancy
infant deaths
90
80
70
60
50
40
0
250
500
750
1000
1250
1500
1750


Scatter plot of Alcohol vs Life expectancy
Life expectancy
Alcohol
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

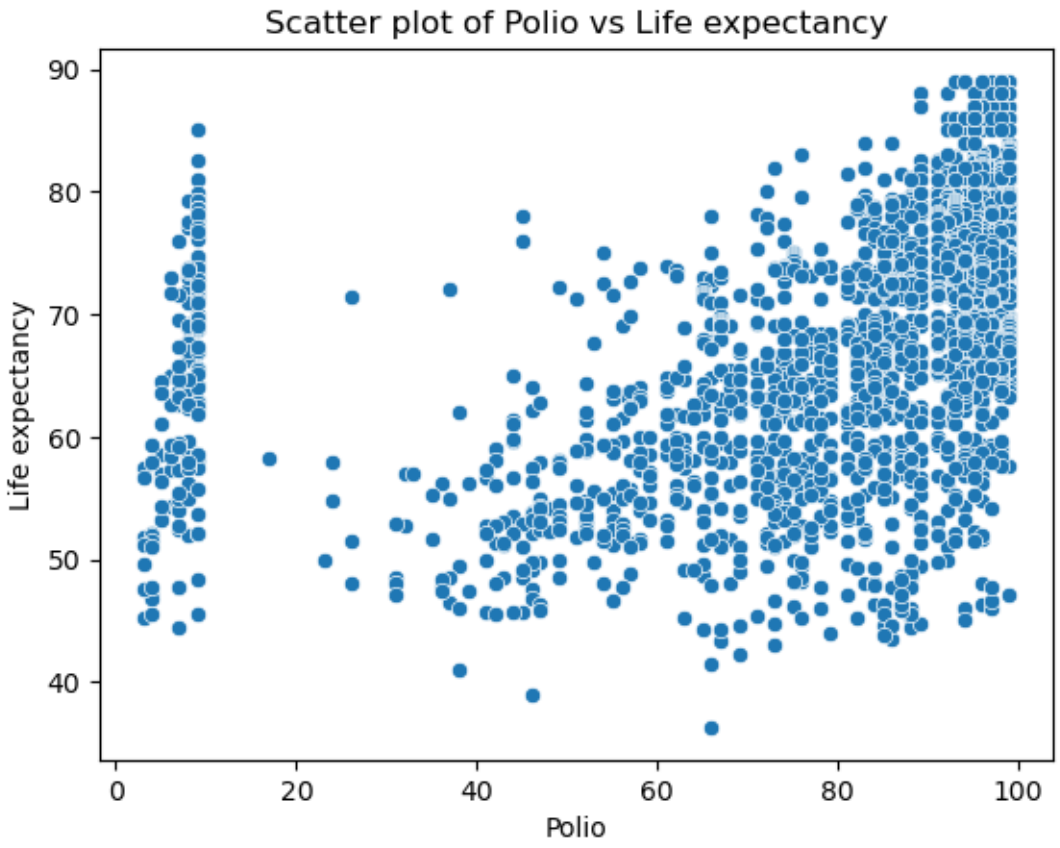

Scatter plot of percentage expenditure vs Life expectancy
Life expectancy
90
80
70
60
50
40
0
2500
5000
7500
10000
12500
15000
17500
20000
percentage expenditure

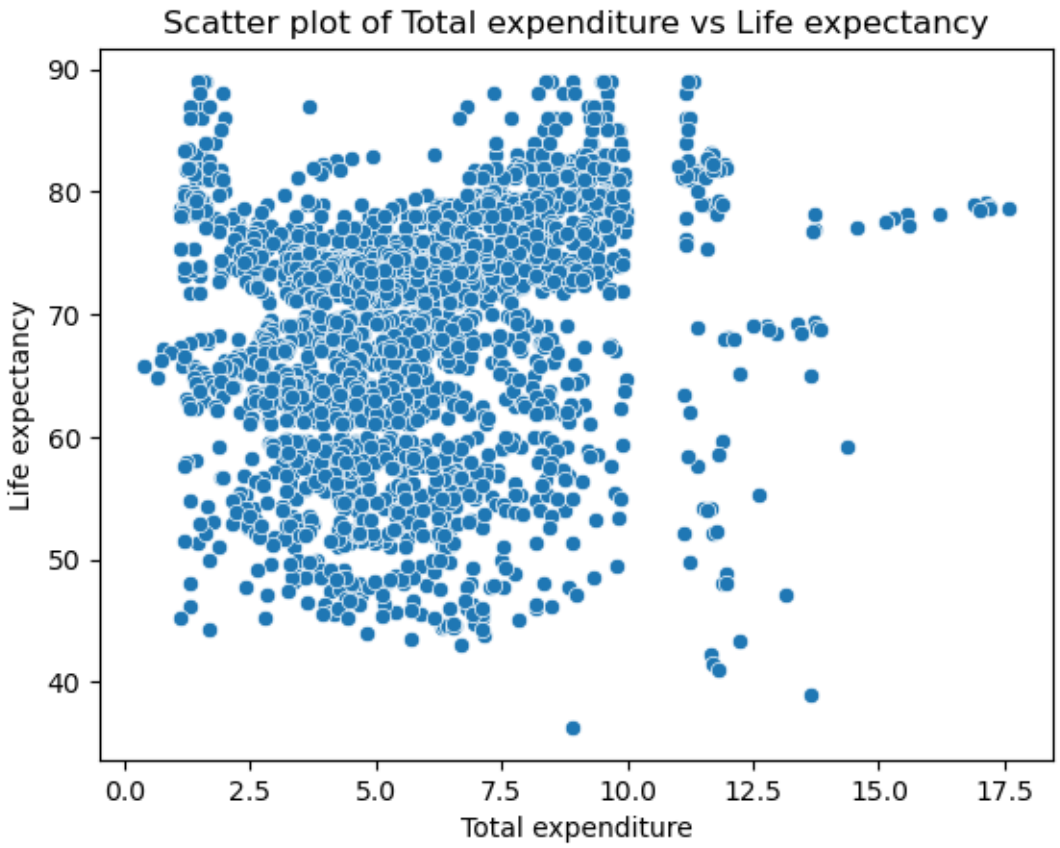

Scatter plot of Hepatitis B vs Life expectancy
Life expectancy
Hepatitis B
90
80
70
60
50
0
20
40
60
80
100

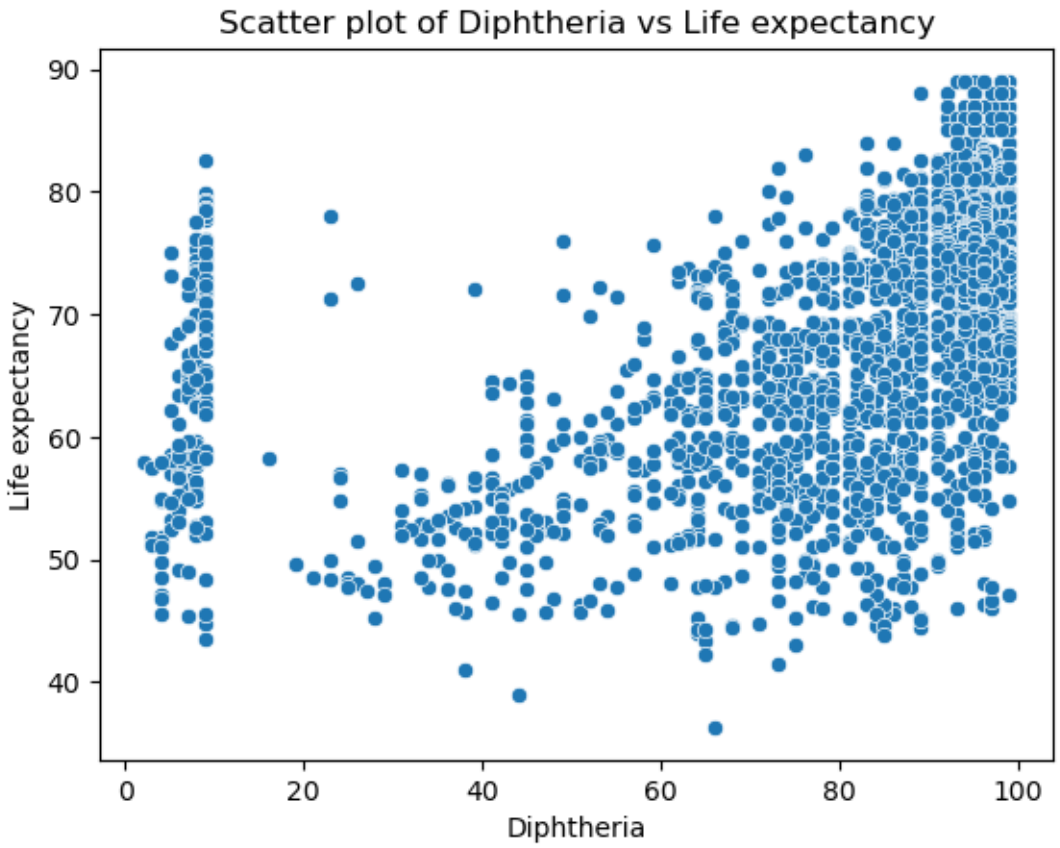

Scatter plot of Measles vs Life expectancy
Life expectancy
90
80
70
60
50
40
0
50000
100000
150000
200000
Measles

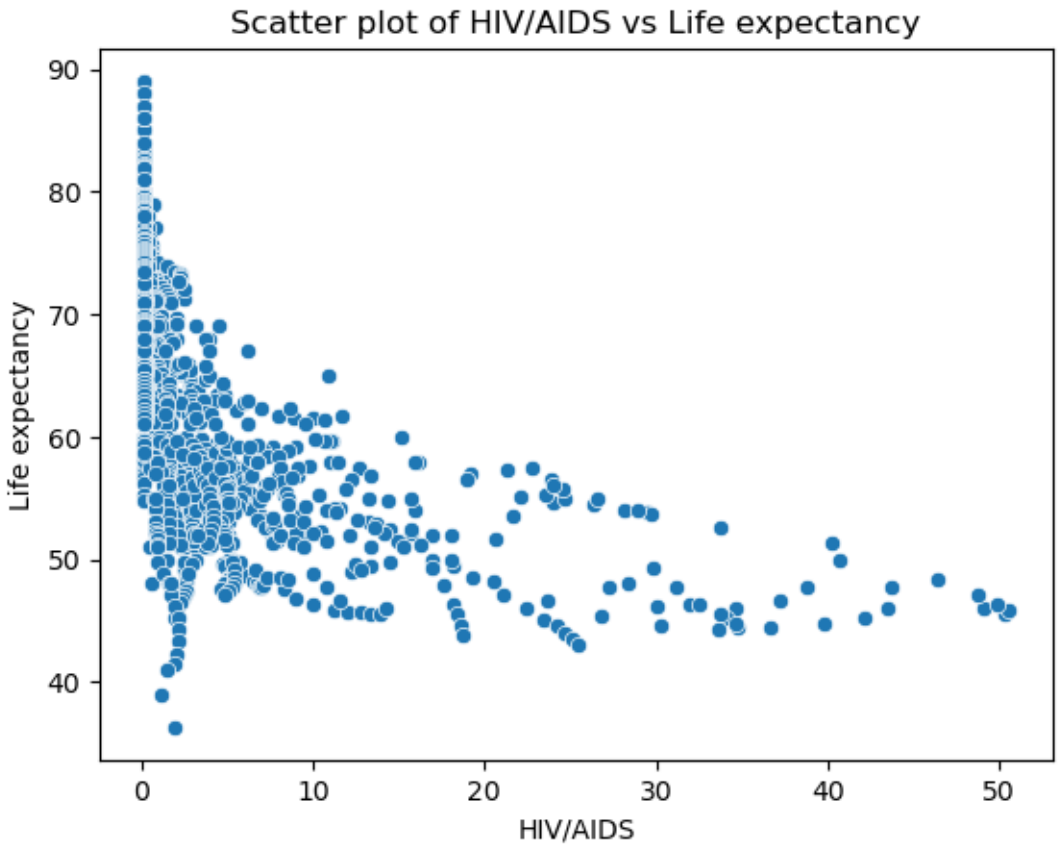

Scatter plot of BMI vs Life expectancy
Life expectancy
BMI
90
80
70
60
50
40
0
10
20
30
40
50
60
70
80


Scatter plot of under-five deaths vs Life expectancy
Life expectancy
under-five deaths
90
80
70
60
50
40
0
500
1000
1500
2000
2500

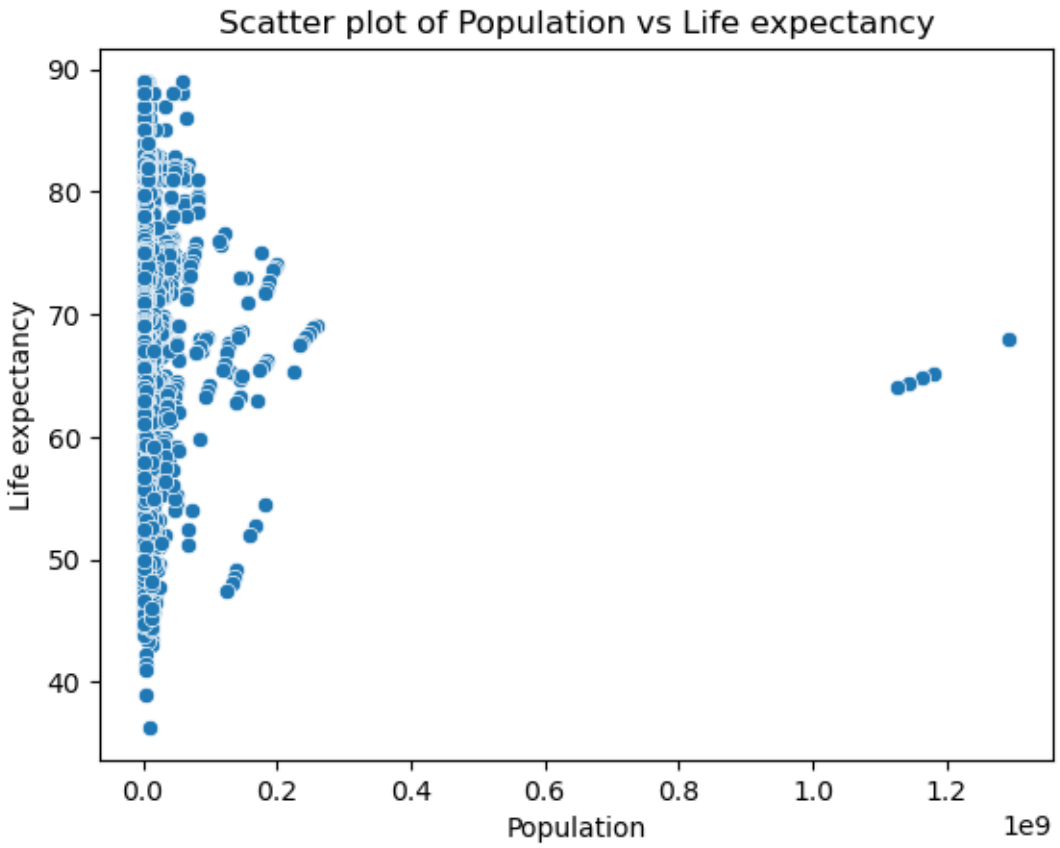

Scatter plot of Polio vs Life expectancy
Life expectancy
Polio
90
80
70
60
50
40
0
20
40
60
80
100

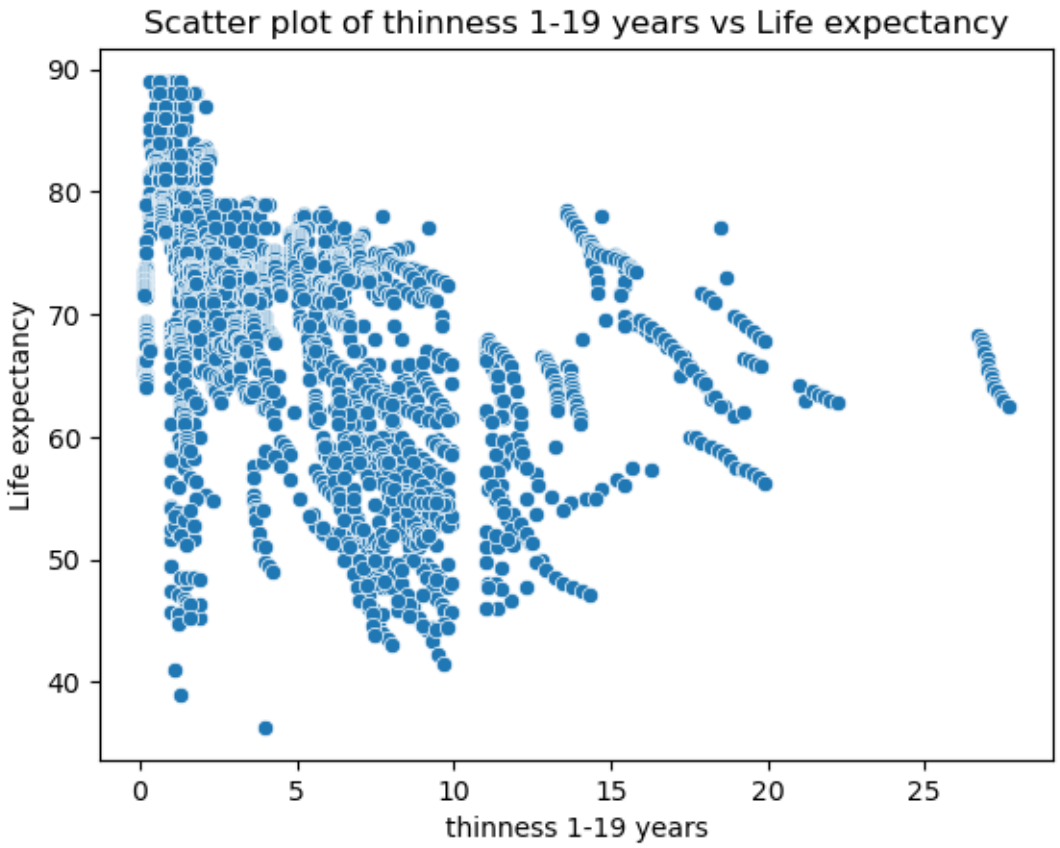

Scatter plot of Total expenditure vs Life expectancy
Life expectancy
Total expenditure
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

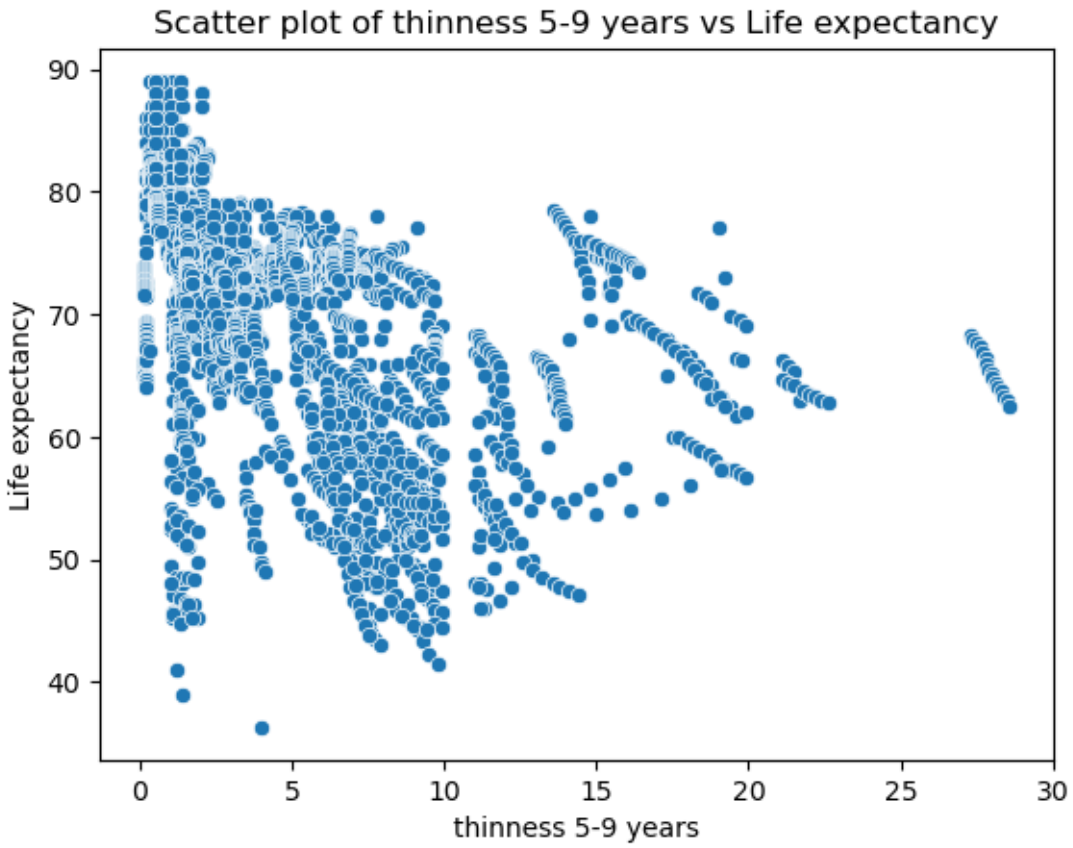

Scatter plot of Diphtheria vs Life expectancy
Life expectancy
Diphtheria
90
80
70
60
50
40
0
20
40
60
80
100


Scatter plot of HIV/AIDS vs Life expectancy
Life expectancy
HIV/AIDS
90
80
70
60
50
40
0
10
20
30
40
50

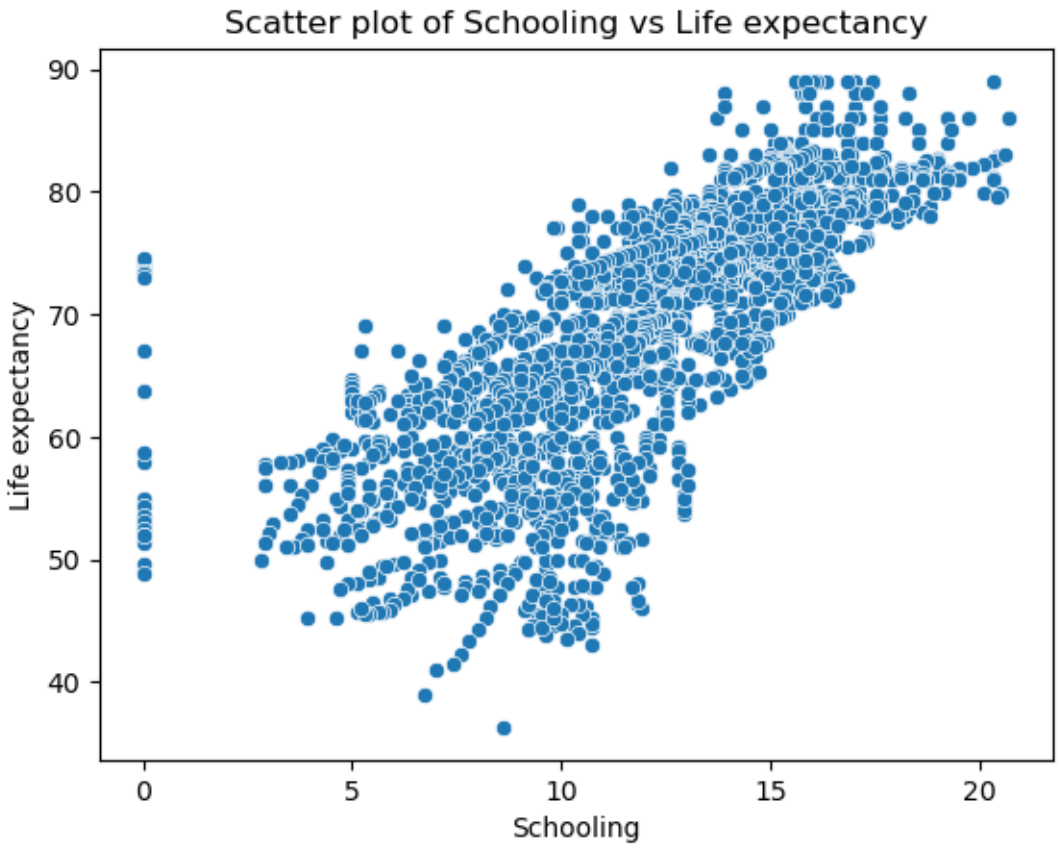

Scatter plot of GDP vs Life expectancy
Life expectancy
GDP
90
80
70
60
50
40
0
20000
40000
60000
80000
100000
120000


Scatter plot of Population vs Life expectancy
Life expectancy
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Population
1e9


Scatter plot of thinness 1-19 years vs Life expectancy
Life expectancy
thinness 1-19 years
90
80
70
60
50
40
0
5
10
15
20
25


Scatter plot of thinness 5-9 years vs Life expectancy
Life expectancy
thinness 5-9 years
90
80
70
60
50
40
0
5
10
15
20
25
30


Scatter plot of Income composition of resources vs Life expectancy
Life expectancy
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8
Income composition of resources



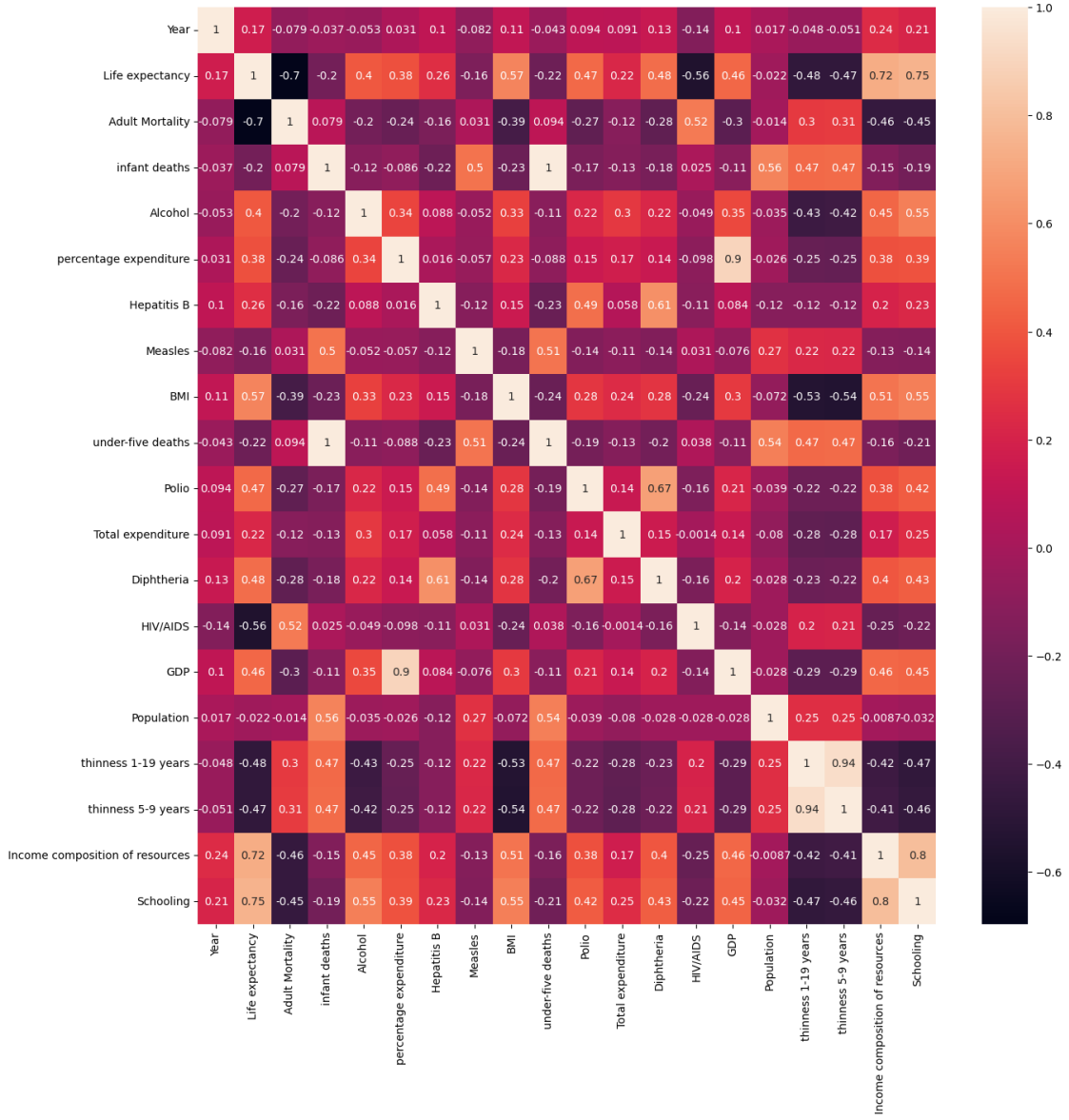
[23]:
```
data.select_dtypes(include='number').columns
```

[23]:
```
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')
```

[54]:
```
## corelation of heatmap
s=data.select_dtypes(include="number").corr()
```

[59]:
```
plt.figure(figsize=(15, 15))
sns.heatmap(s,annot=True)
```

[59]:
```
<Axes: >
```



2 Missing values treatments

[60]:
```
data.isnull().sum()
```

[60]:
```
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
```

[61]:
```
for i in ["BMI","Polio","Income composition of resources"]:
    data[i].fillna(data[i].median(),inplace=True)
```

```
C:\Users\Vikas\AppData\Local\Temp\ipykernel_23700\3528658117.py:2:
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  data[i].fillna(data[i].median(),inplace=True)
C:\Users\Vikas\AppData\Local\Temp\ipykernel_23700\3528658117.py:2:
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  data[i].fillna(data[i].median(),inplace=True)
C:\Users\Vikas\AppData\Local\Temp\ipykernel_23700\3528658117.py:2:
```

```
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  data[i].fillna(data[i].median(),inplace=True)
```

[62]:
```
data.isnull().sum()
```

[62]:
```
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                                0
under-five deaths                  0
Polio                              0
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    0
Schooling                        163
dtype: int64
```

[64]:
```
## remove the missing values knnImputabser
from sklearn.impute import KNNImputer
impute=KNNImputer()
```

[65]:
```
for i in data.select_dtypes(include="number").columns:
    data[i]=impute.fit_transform(data[[i]])
```

[66]:
```
# ofter we check the null values
data.isnull().sum()
```

[66]:
```
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
```

3 Outliers treatment

[77]:
```
## Decide whether to do outlier treatment or not, if do how?  outlier treatment␣
 ↪we do the continues numerical columns
def wisker(col):
    q1, q3 = np.percentile(col, [25, 75])  # Calculate the first and third␣
 ↪quartile
    iqr = q3 - q1  # Calculate the interquartile range (IQR)
    lw = q1 - 1.5 * iqr  # Lower whisker
    uw = q3 + 1.5 * iqr  # Upper whisker
    return lw, uw
```

[78]:
```
data.columns
```

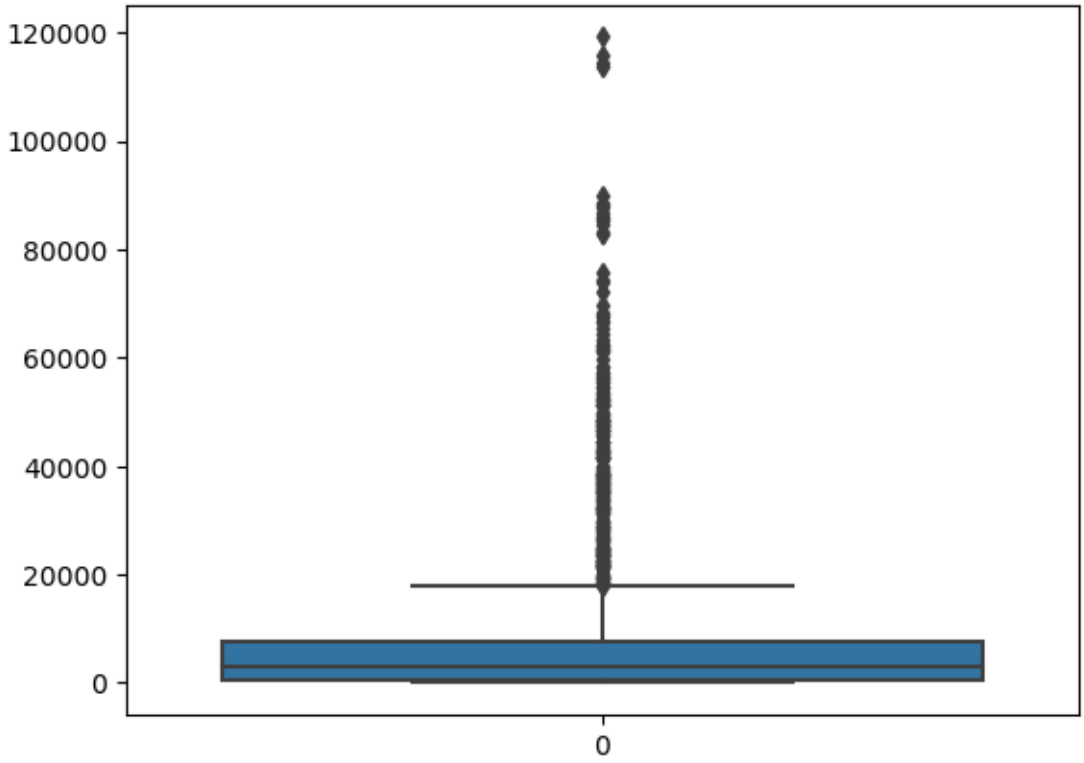
[78]:
```
Index(['Country', 'Year', 'Status', 'Life expectancy', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B',
       'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total expenditure',
       'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness 1-19 years',
       'thinness 5-9 years', 'Income composition of resources', 'Schooling'],
      dtype='object')
```

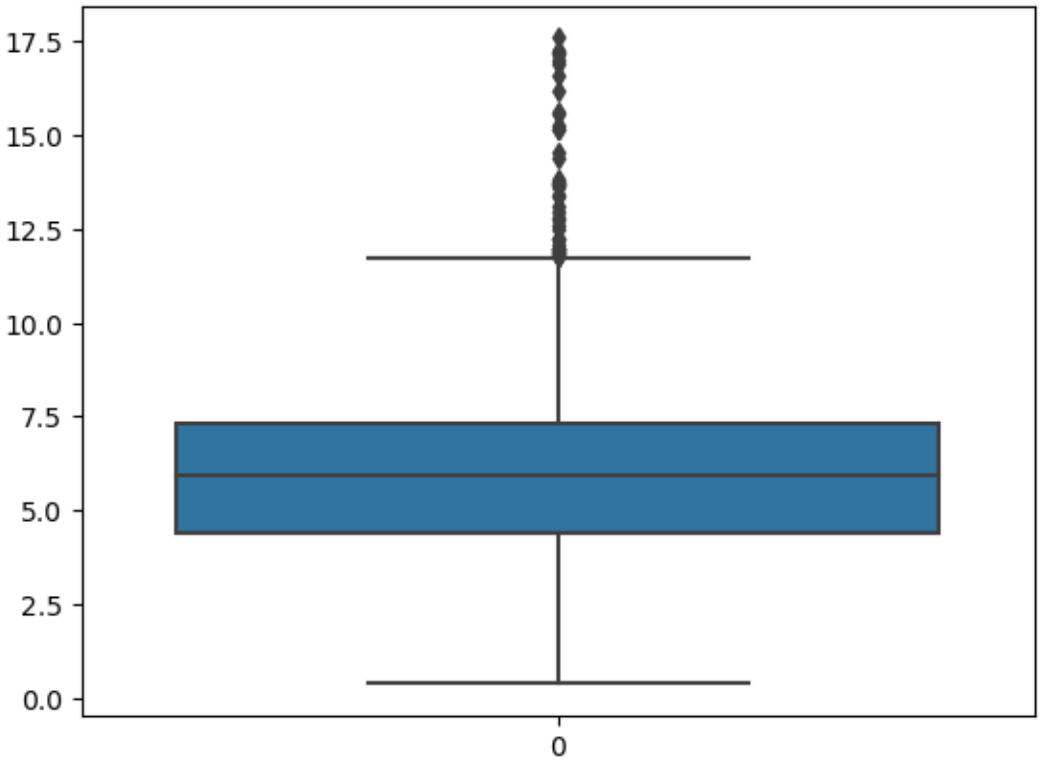
[80]:
```
wisker(data['GDP'])
```

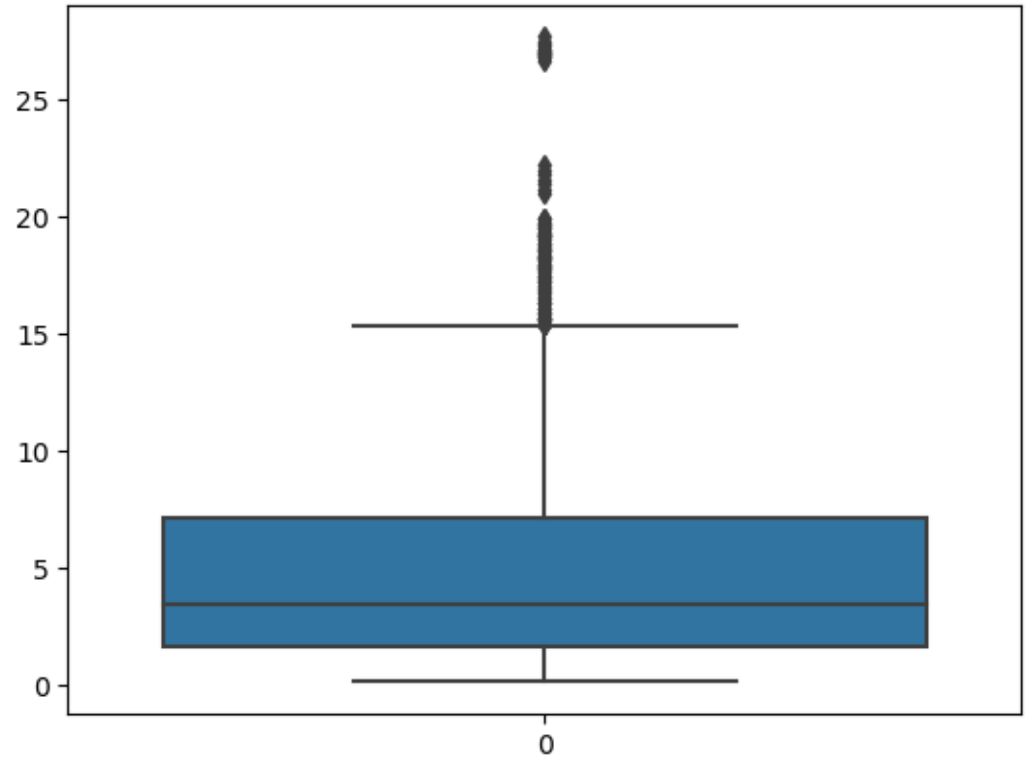
[80]: (-9773.52021495771, 17837.165679596183)

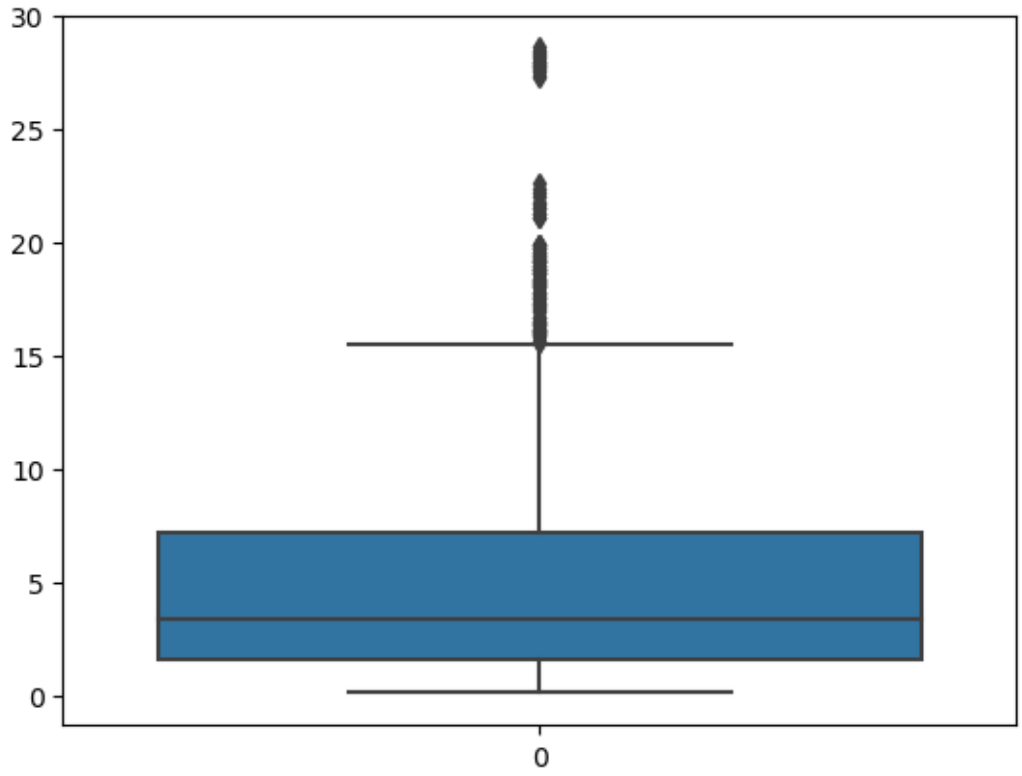
[ ]:
```
for i in ['GDP','Total expenditure','thinness 1-19 years','thinness 5-9 years']:
    lw,uw=wisker(data[i])
    data[i]=np.where(data[i]<lw,lw,data[i])
    data[i]=np.where(data[i]>uw,uw,data[i])
```

[81]:
```
for i in ['GDP','Total expenditure','thinness 1-19 years','thinness 5-9 years']:
    sns.boxplot(data[i])
    plt.show()
```




17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0


25
20
15
10
5
0
0



[83]:
```
dummy=pd.get_dummies(data=data,columns=["Country","Status"],drop_first=True)
```

[84]:
```
dummy
```

[84]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	\
0	2015.0	65.0	263.0	62.0	0.01	
1	2014.0	59.9	271.0	64.0	0.01	
2	2013.0	59.9	268.0	66.0	0.01	
3	2012.0	59.5	272.0	69.0	0.01	
4	2011.0	59.2	275.0	71.0	0.01	
…	…	…	…	…	…	
2933	2004.0	44.3	723.0	27.0	4.36	
2934	2003.0	44.5	715.0	26.0	4.06	
2935	2002.0	44.8	73.0	25.0	4.43	
2936	2001.0	45.3	686.0	25.0	1.72	
2937	2000.0	46.0	665.0	24.0	1.68	

	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	\
0	71.279624	65.0	1154.0	19.1	83.0	

```
1                  73.523582         62.0    492.0  18.6              86.0
2                  73.219243         64.0    430.0  18.1              89.0
3                  78.184215         67.0   2787.0  17.6              93.0
4                   7.097109         68.0   3013.0  17.2              97.0
…                        …          …       …   …                 …
2933               0.000000         68.0     31.0  27.1              42.0
2934               0.000000          7.0    998.0  26.7              41.0
2935               0.000000         73.0    304.0  26.3              40.0
2936               0.000000         76.0    529.0  25.9              39.0
2937               0.000000         79.0   1483.0  25.5              39.0

      …  Country_United States of America  Country_Uruguay  \
0     …                              False            False
1     …                              False            False
2     …                              False            False
3     …                              False            False
4     …                              False            False
…   …                                …                …
2933  …                              False            False
2934  …                              False            False
2935  …                              False            False
2936  …                              False            False
2937  …                              False            False

      Country_Uzbekistan  Country_Vanuatu  \
0                  False            False
1                  False            False
2                  False            False
3                  False            False
4                  False            False
…                  …                …
2933               False            False
2934               False            False
2935               False            False
2936               False            False
2937               False            False

      Country_Venezuela (Bolivarian Republic of)  Country_Viet Nam  \
0                                          False             False
1                                          False             False
2                                          False             False
3                                          False             False
4                                          False             False
…                                          …                 …
2933                                       False             False
2934                                       False             False
2935                                       False             False
```

```
2936                                    False             False
2937                                    False             False

      Country_Yemen  Country_Zambia  Country_Zimbabwe  Status_Developing
0             False           False              False               True
1             False           False              False               True
2             False           False              False               True
3             False           False              False               True
4             False           False              False               True
…               …               …                  …                  …
2933          False           False               True               True
2934          False           False               True               True
2935          False           False               True               True
2936          False           False               True               True
2937          False           False               True               True

[2938 rows x 213 columns]
```

[ ]: