IMPORT THE LIBRARIES

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
```

IMPORT THE DATASET

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

Out[2]:

| | Country | Year | Status | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentag expenditu |
|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2015 | Developing | 65.0 | 263.0 | 62 | 0.01 | 71.2796 |
| **1** | Afghanistan | 2014 | Developing | 59.9 | 271.0 | 64 | 0.01 | 73.5235 |
| **2** | Afghanistan | 2013 | Developing | 59.9 | 268.0 | 66 | 0.01 | 73.2192 |
| **3** | Afghanistan | 2012 | Developing | 59.5 | 272.0 | 69 | 0.01 | 78.1842 |
| **4** | Afghanistan | 2011 | Developing | 59.2 | 275.0 | 71 | 0.01 | 7.0971 |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **2933** | Zimbabwe | 2004 | Developing | 44.3 | 723.0 | 27 | 4.36 | 0.0000 |
| **2934** | Zimbabwe | 2003 | Developing | 44.5 | 715.0 | 26 | 4.06 | 0.0000 |
| **2935** | Zimbabwe | 2002 | Developing | 44.8 | 73.0 | 25 | 4.43 | 0.0000 |
| **2936** | Zimbabwe | 2001 | Developing | 45.3 | 686.0 | 25 | 1.72 | 0.0000 |
| **2937** | Zimbabwe | 2000 | Developing | 46.0 | 665.0 | 24 | 1.68 | 0.0000 |

2938 rows × 22 columns

DATA EXPLORATION

In [3]:
```
df.head(5)
```

Out[3]:

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

5 rows × 22 columns

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

Out[4]:
```
(2938, 22)
```

In [5]:
```
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 [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
df.duplicated().sum()
```

Out[7]:
```
0
```

In [8]:
```
df.describe()
```

Out[8]:

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

HANDLING MISSING VALUES

In [9]:
```
df.fillna(df.median(numeric_only=True),inplace=True)
```

In [10]:
```
categorical_cols=df.select_dtypes(include=['object']).columns
for col in categorical_cols:
    df[col].fillna(df[col].mode()[0])
```

```
print("\n Missing values after imputation:")
print(df.isnull().sum())
```

```
 Missing values after imputation:
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 [ ]:
```
STANDARDIZING THE FEATURES
```

In [11]:
```
df=pd.get_dummies(df, drop_first=True)
```

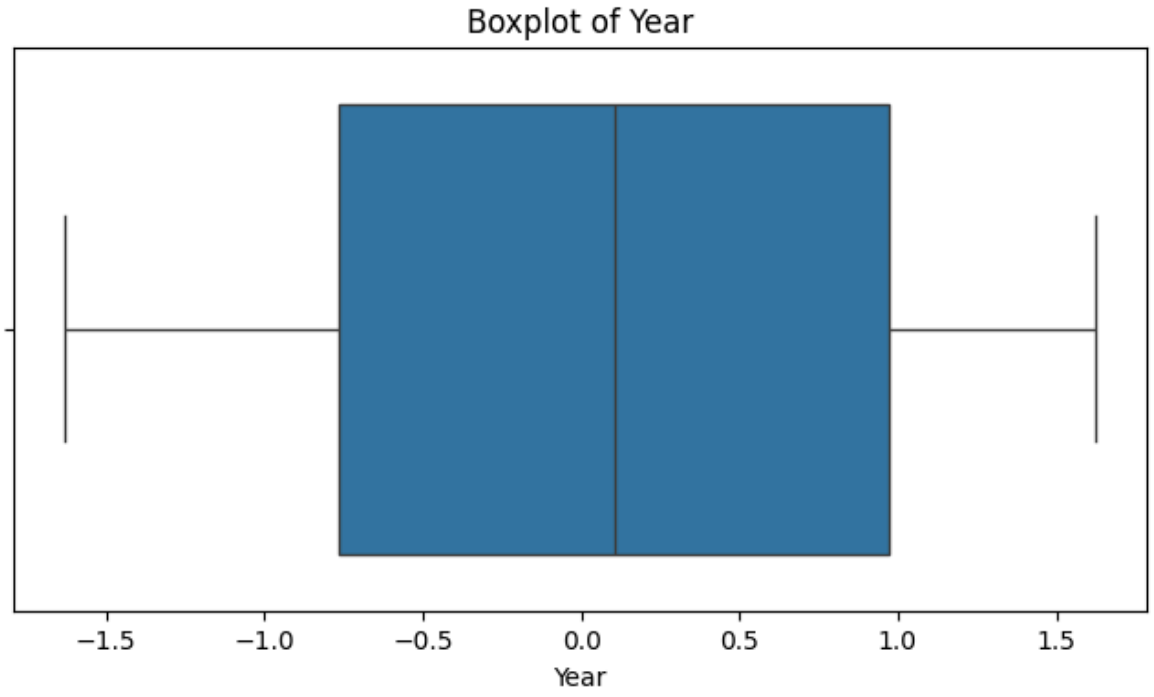
In [12]:
```
numeric_cols=df.select_dtypes(include=['int64','float64']).columns
scaler=StandardScaler()
df[numeric_cols]=scaler.fit_transform(df[numeric_cols])
data_scaled=scaler.transform(df[numeric_cols])
print(data_scaled)
```

```
[[-434.83042485   -7.3289667    -1.32135105 ...   -0.46303607
   -6.67139887   -3.85802803]
 [-434.87741671   -7.38538731   -1.3208313  ...   -0.45309375
   -6.74271209   -3.86741109]
 [-434.92440857   -7.38538731   -1.32102621 ...   -0.44315143
   -6.88533853   -3.87679415]
 ...
 [-435.44131899   -7.55243659   -1.333695   ...   -1.25842143
   -7.90749471   -3.86741109]
 [-435.48831085   -7.54690515   -1.2938695  ...   -1.2385368
   -7.90749471   -3.88617721]
 [-435.53530271   -7.53916115   -1.29523384 ...   -0.76627674
   -7.74109719   -3.88617721]]
```

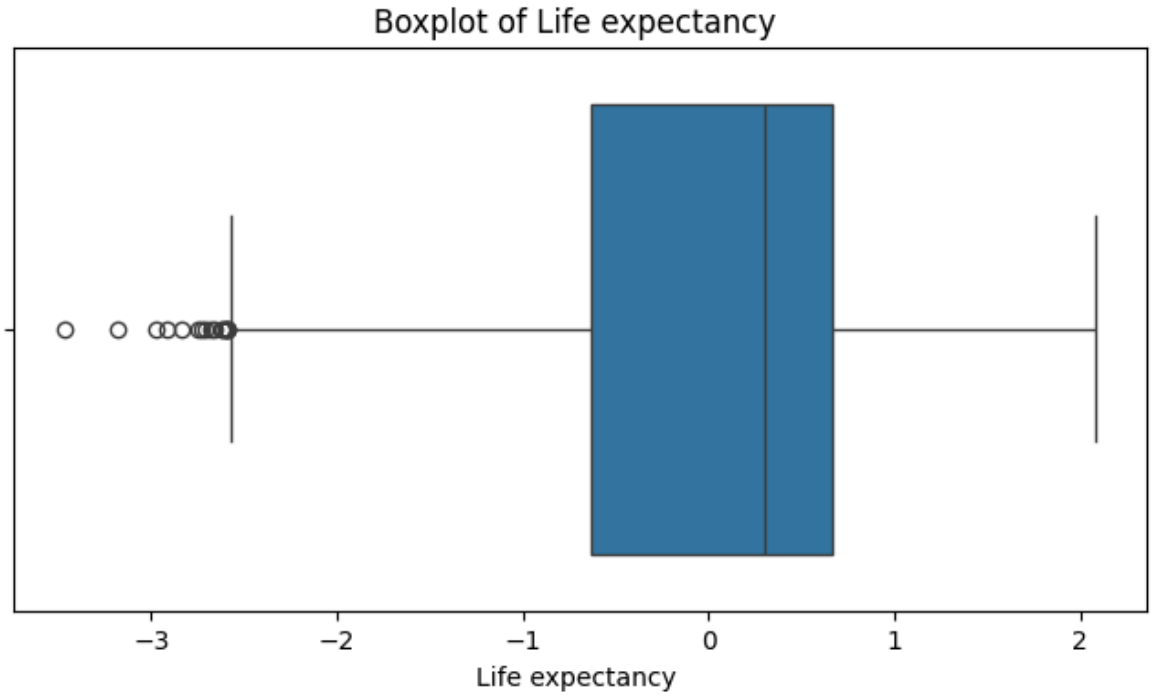
VISUALIZING THE FEATURES

In [13]:
```
for col in numeric_cols:
    plt.figure(figsize=(8,4))
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
```

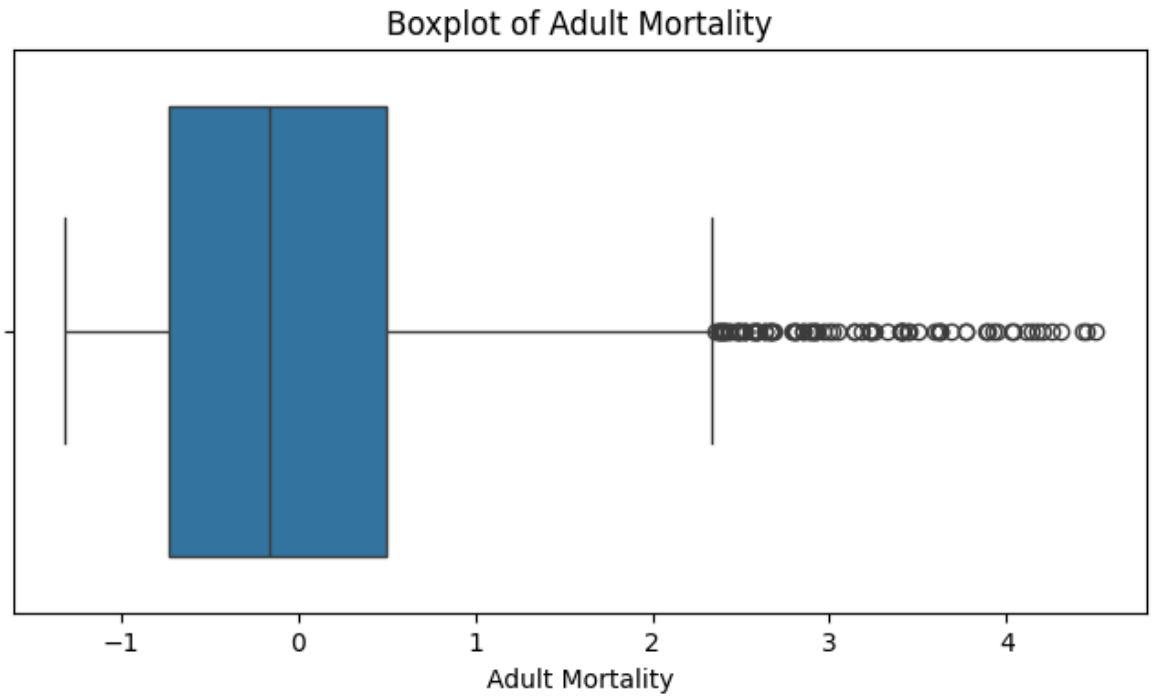
```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



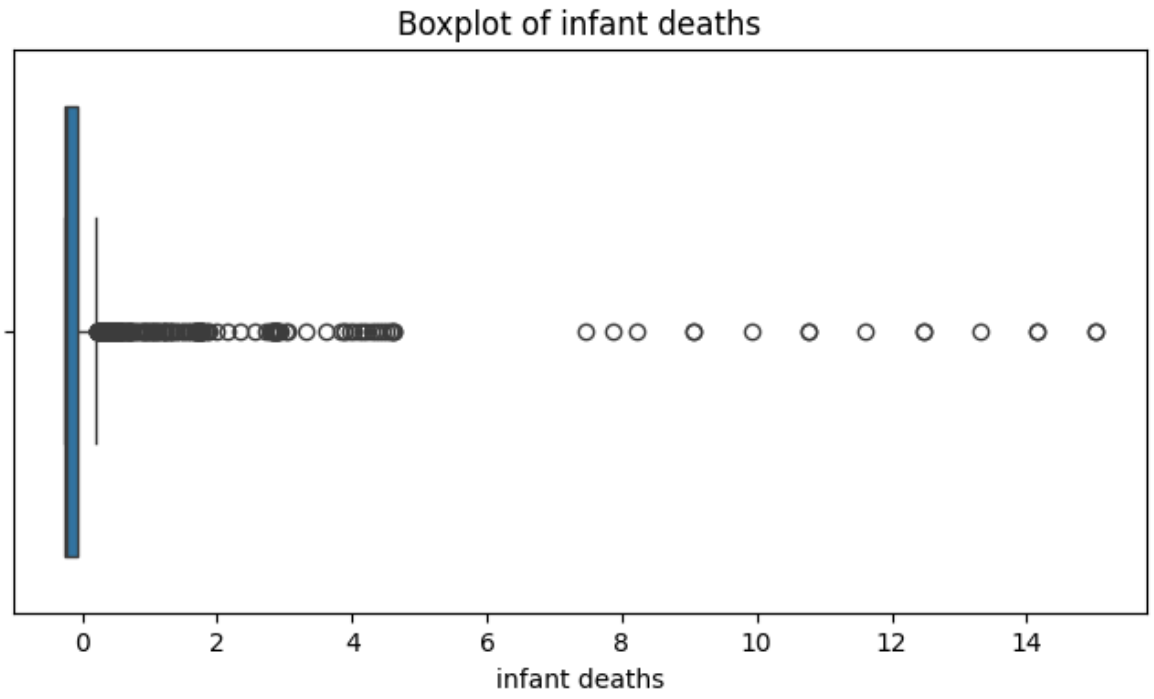

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



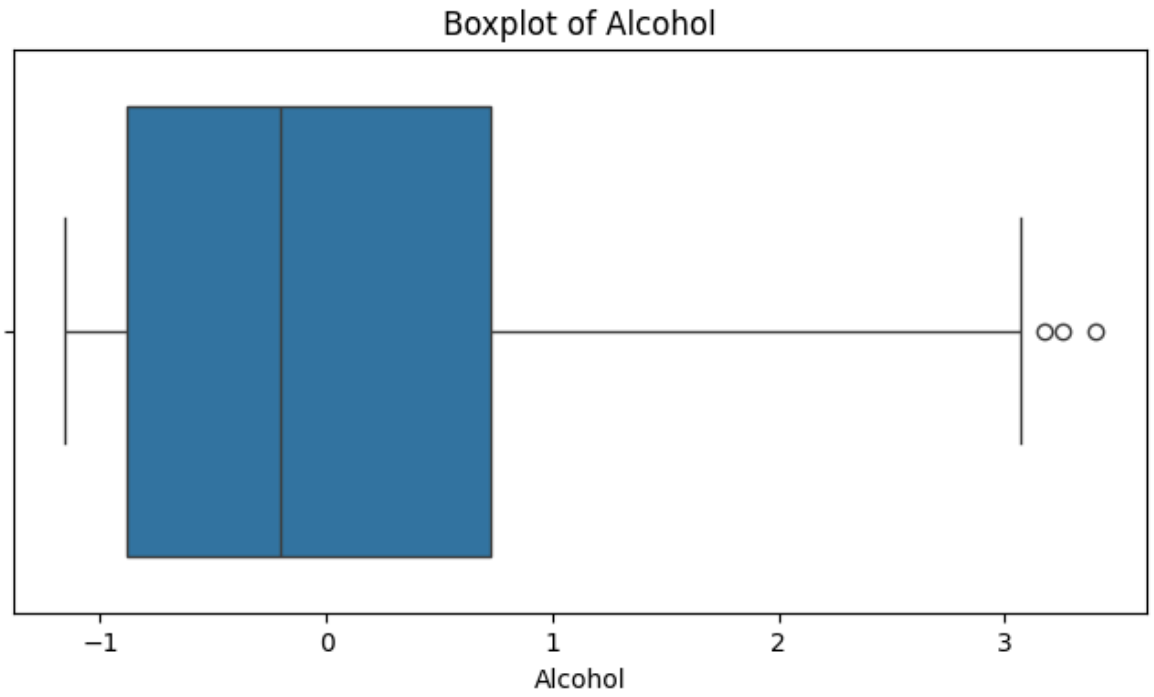

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



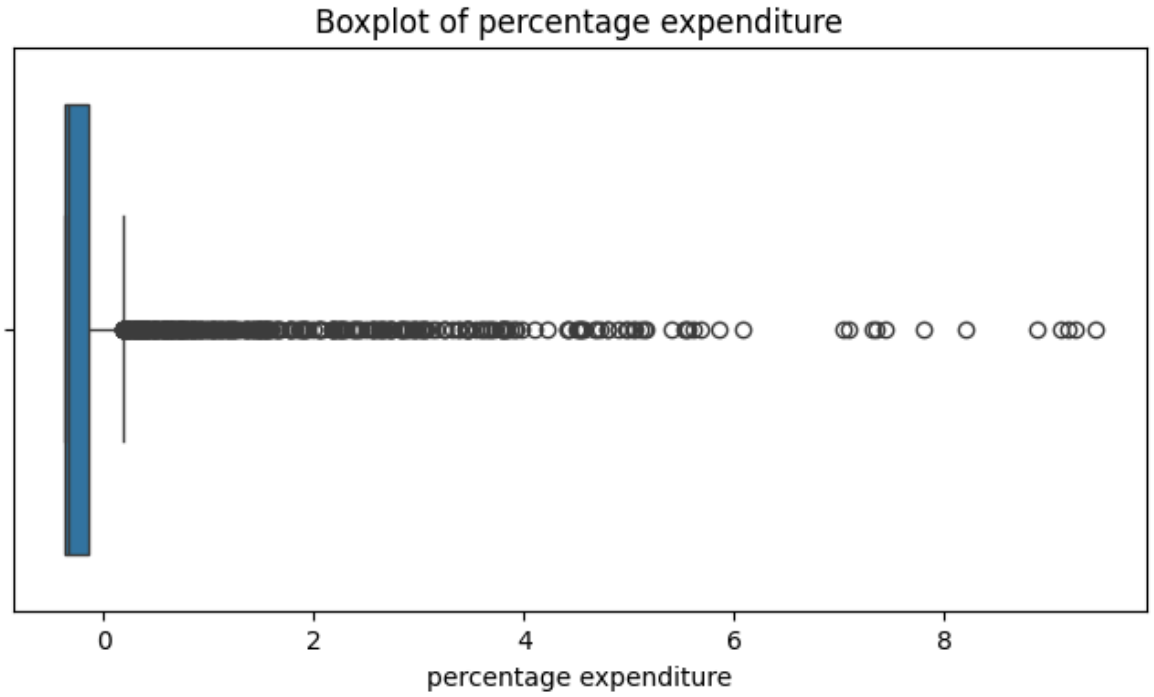

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



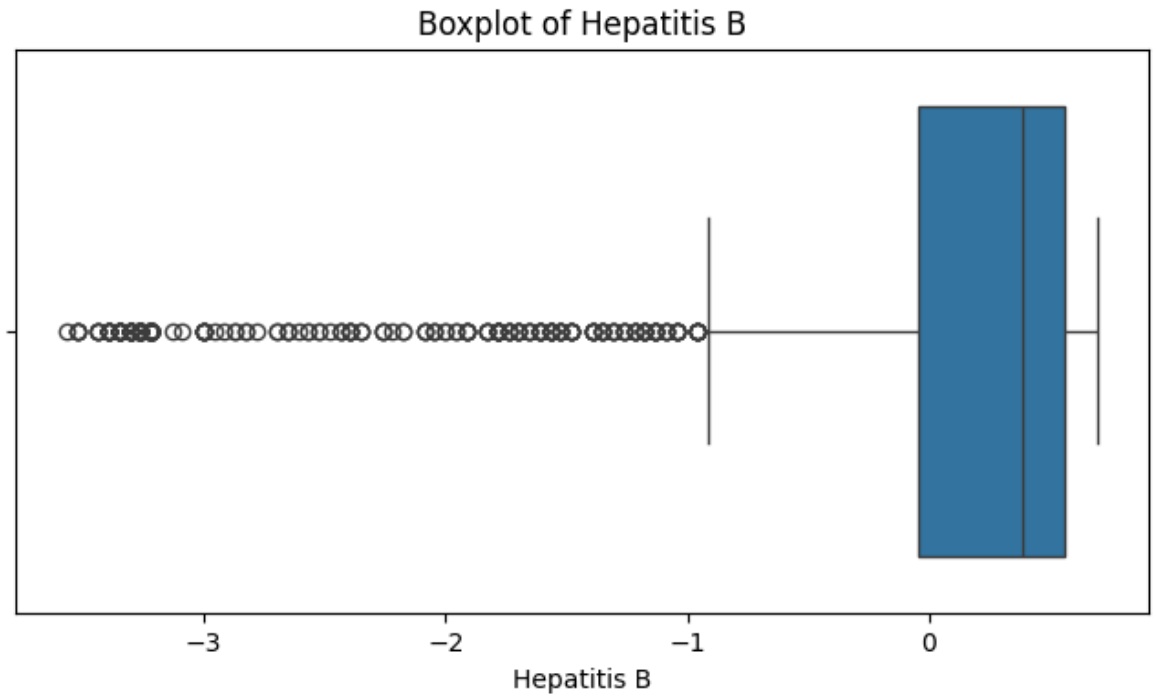

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



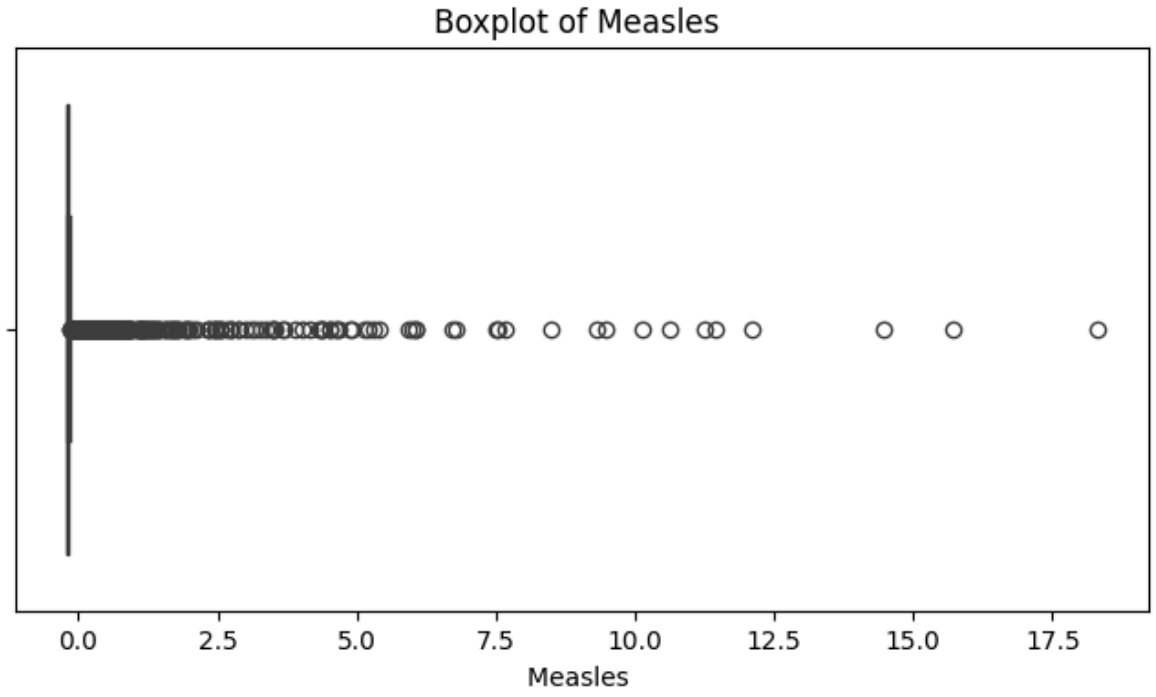

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



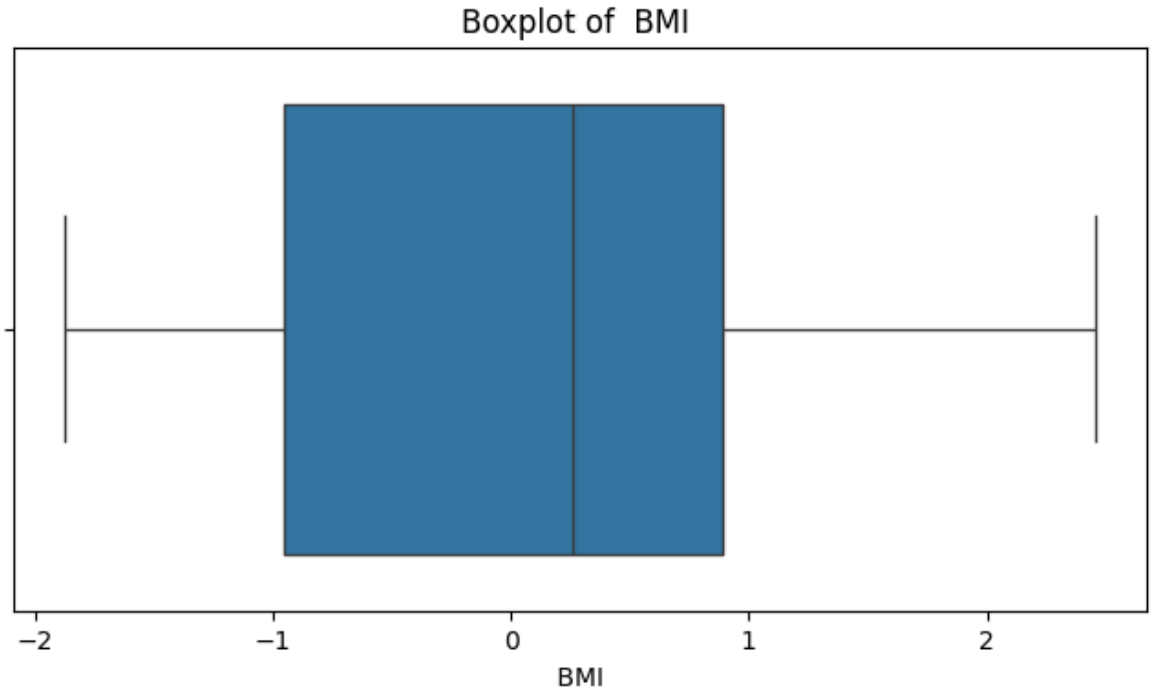

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



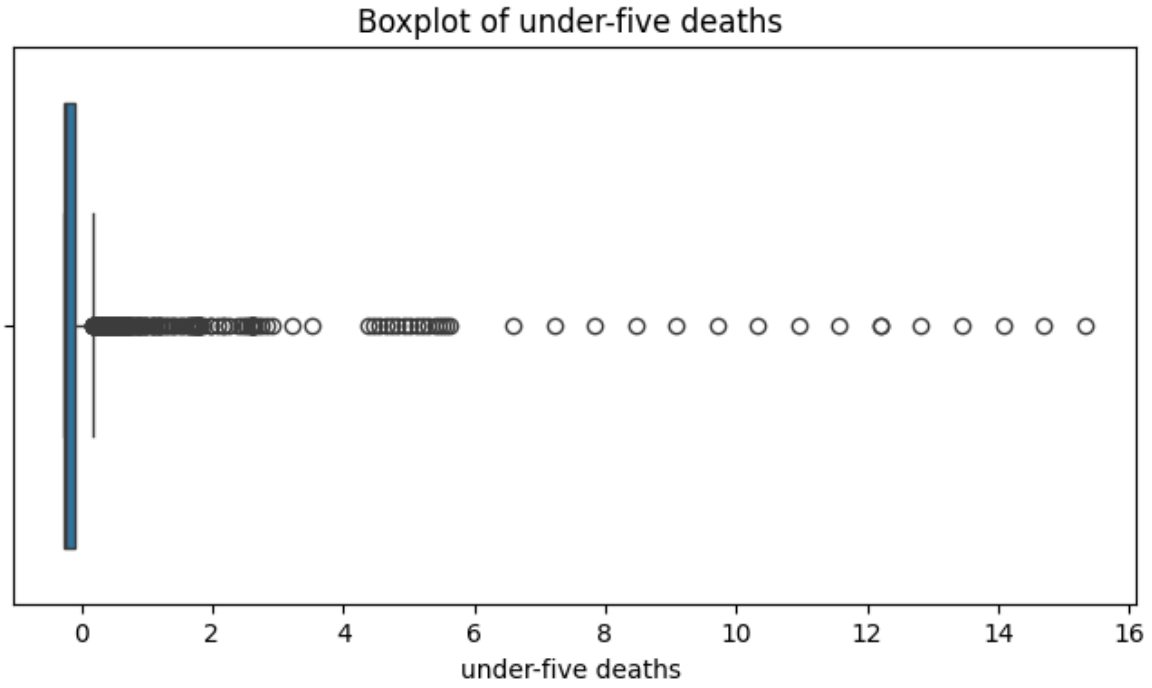

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



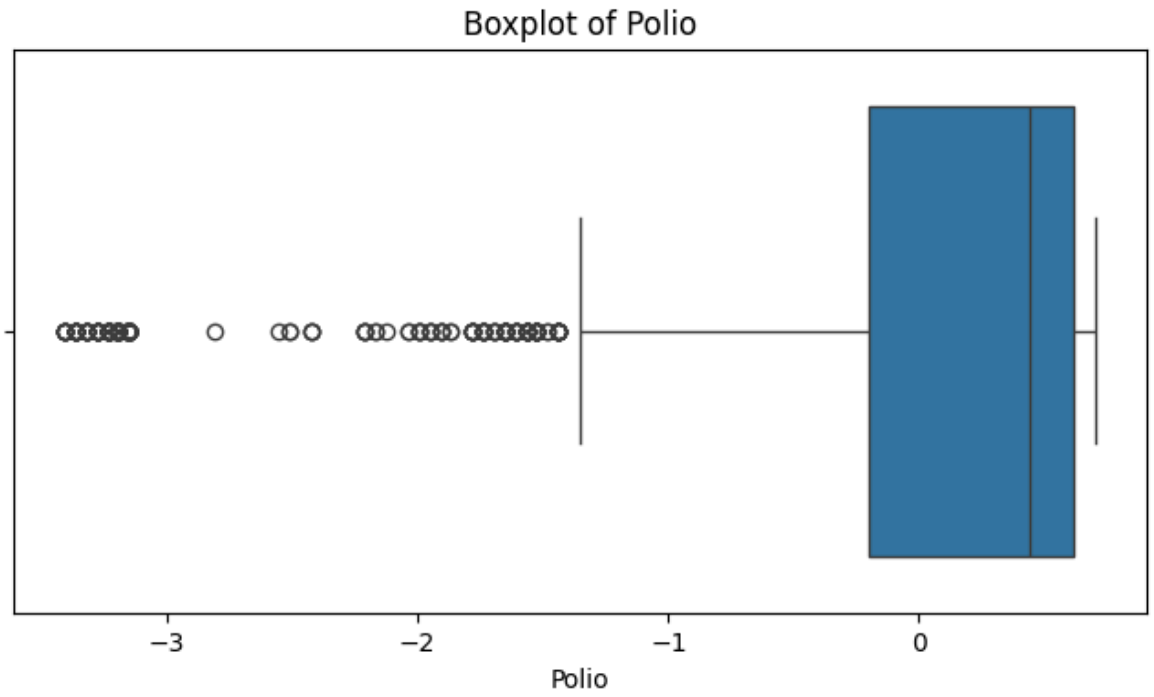

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



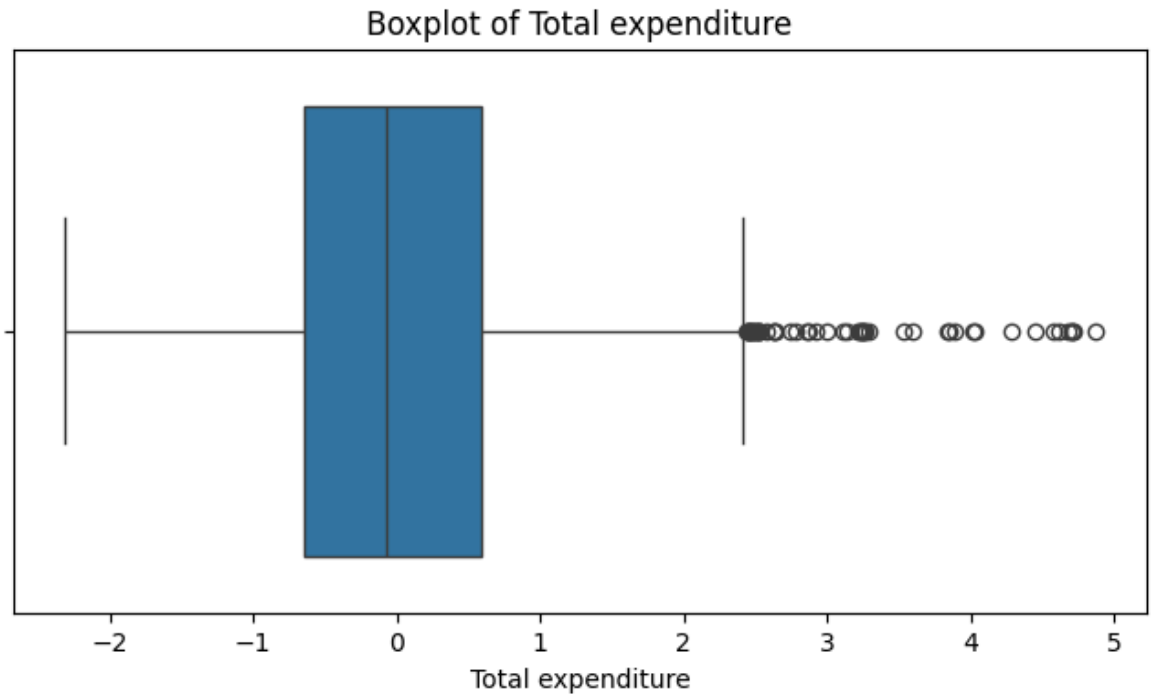

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



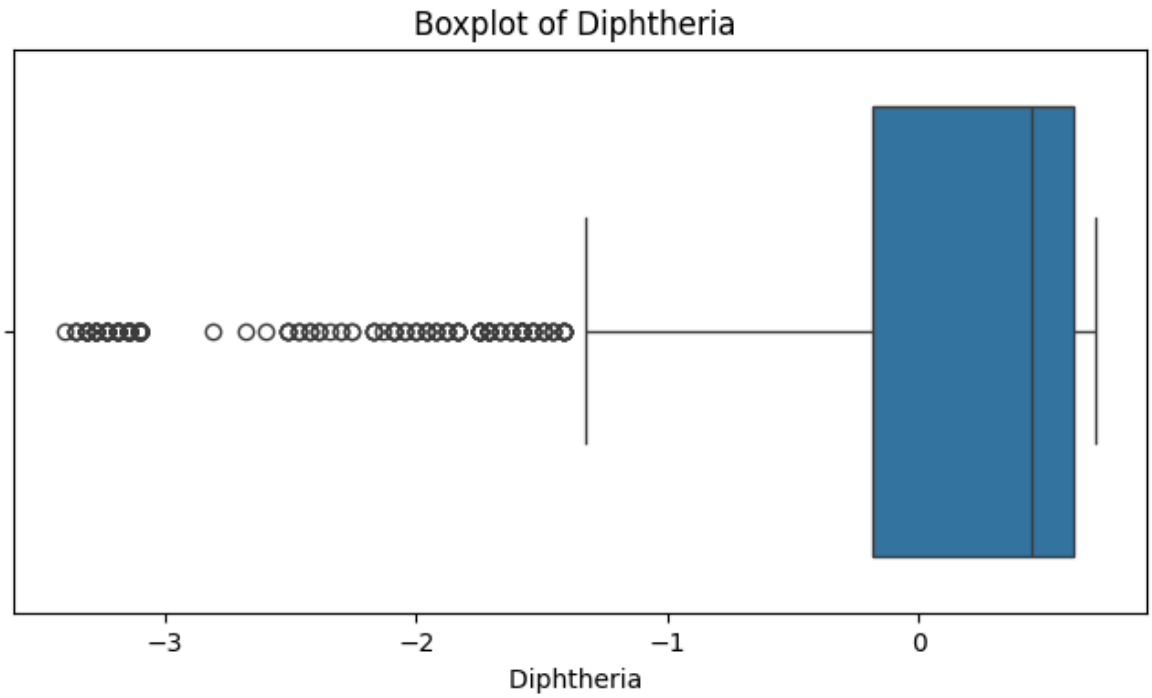

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



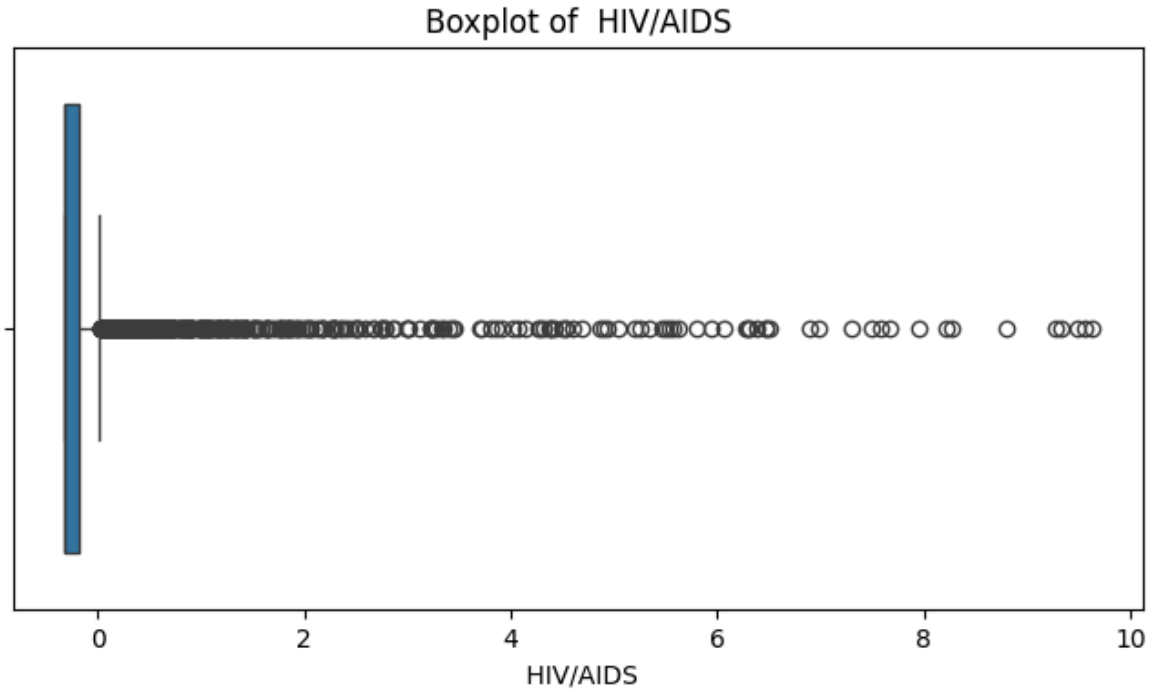

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



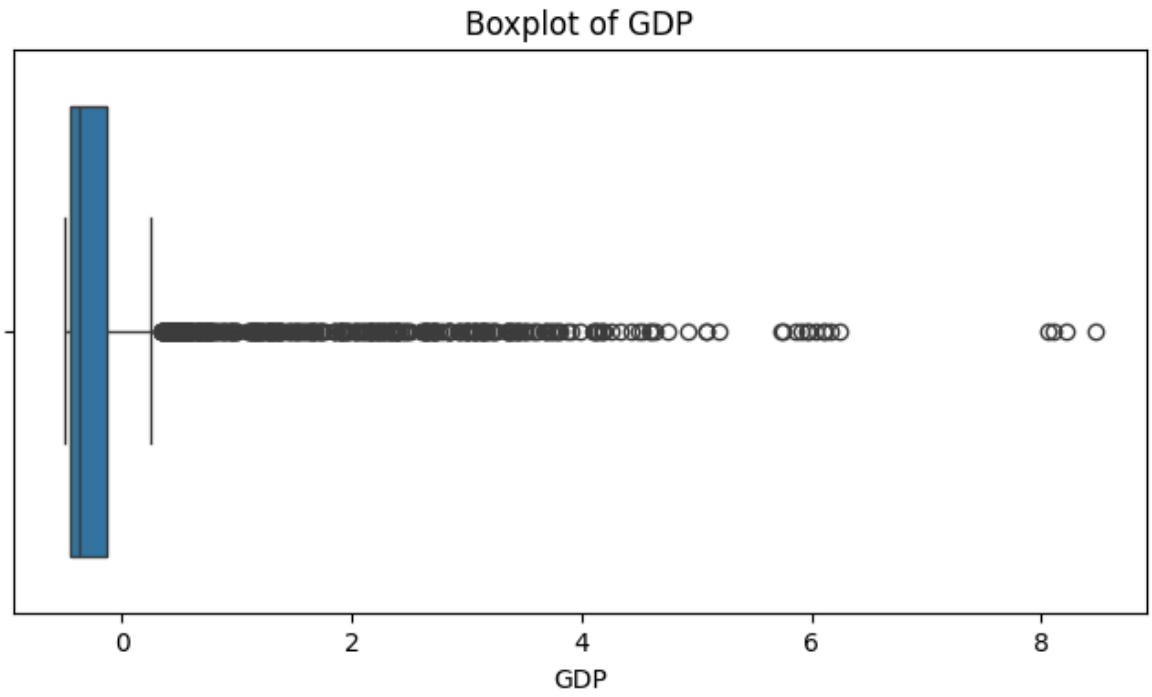

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



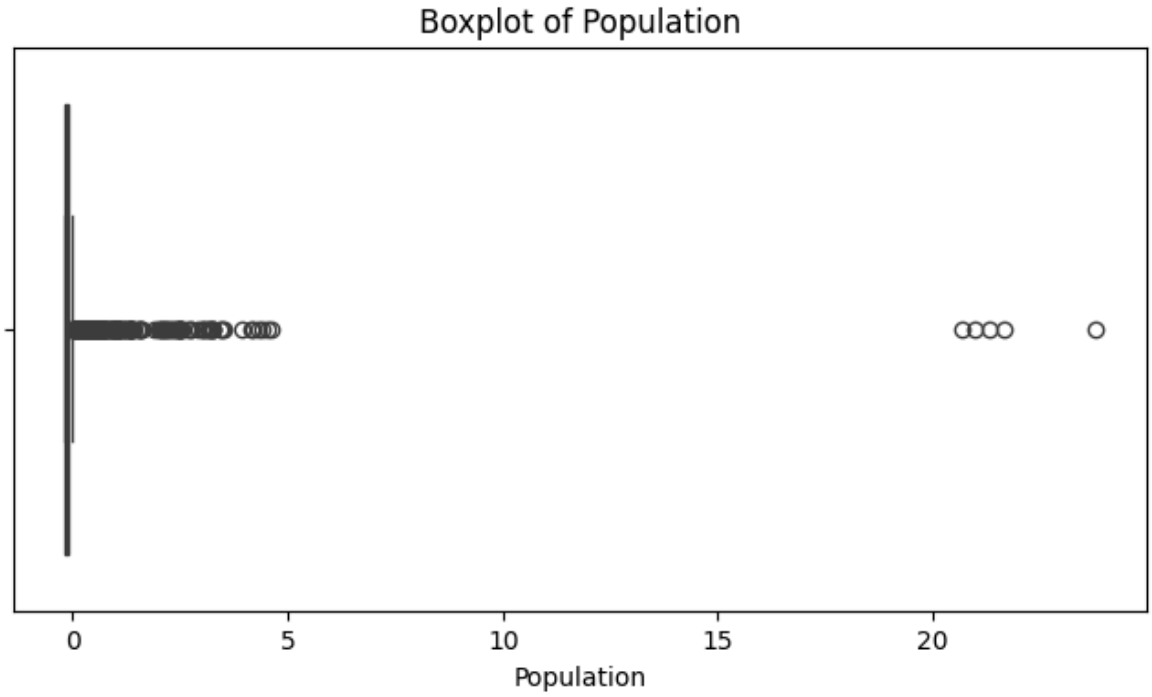

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



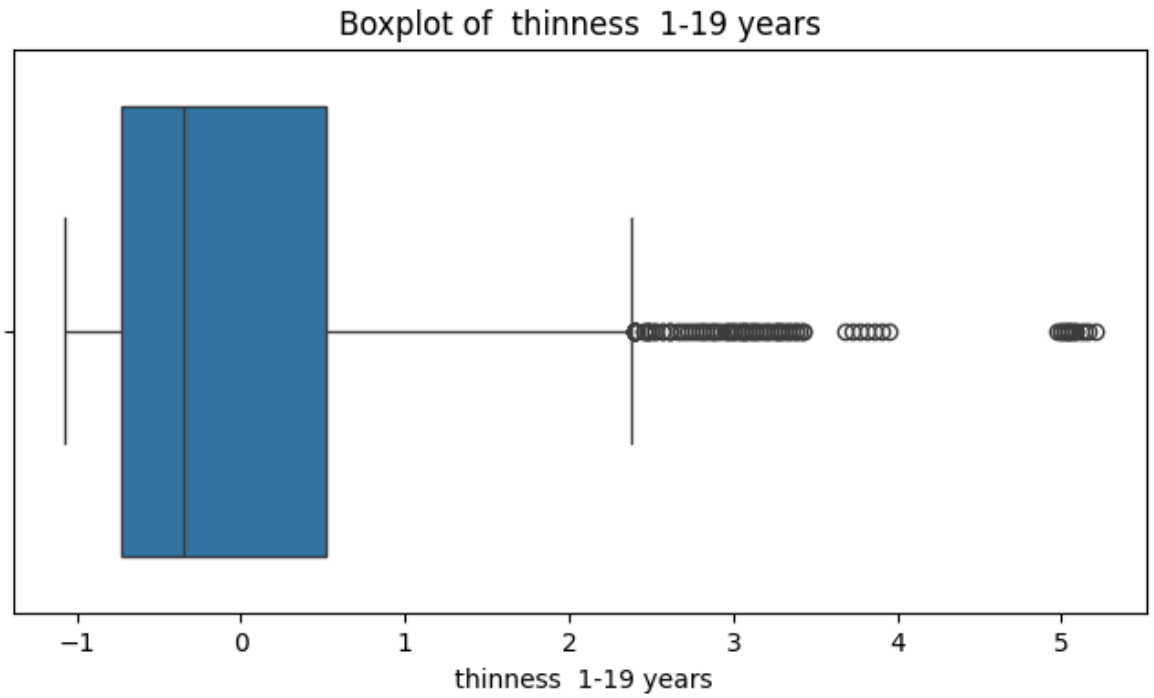

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



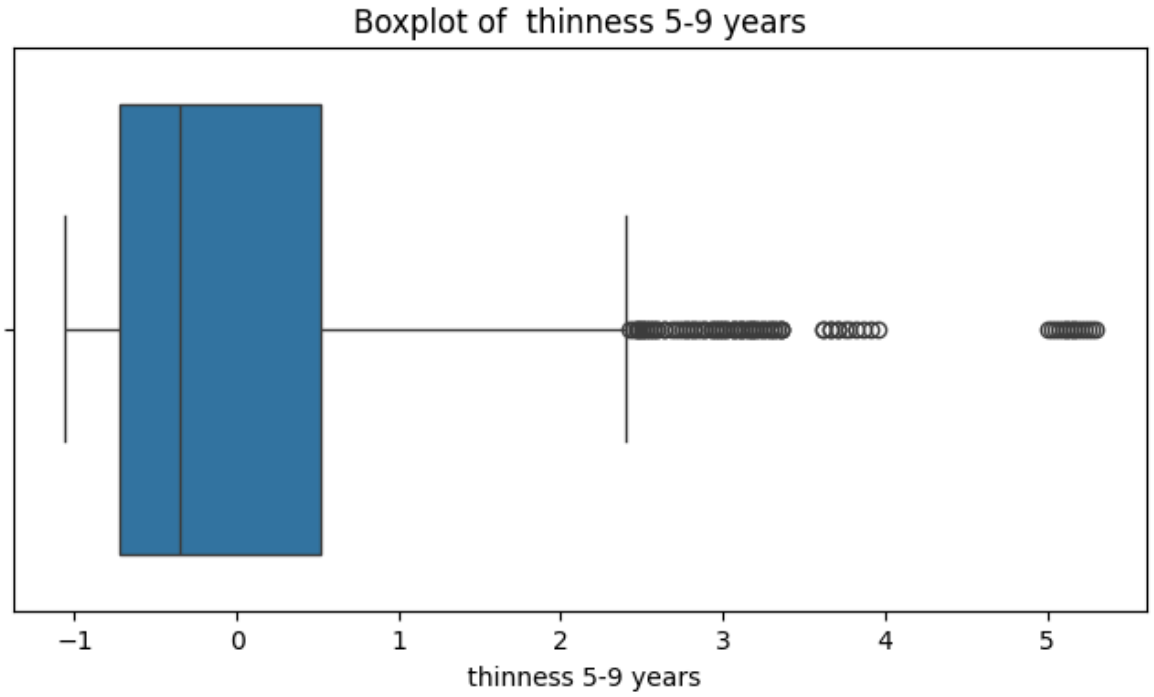

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



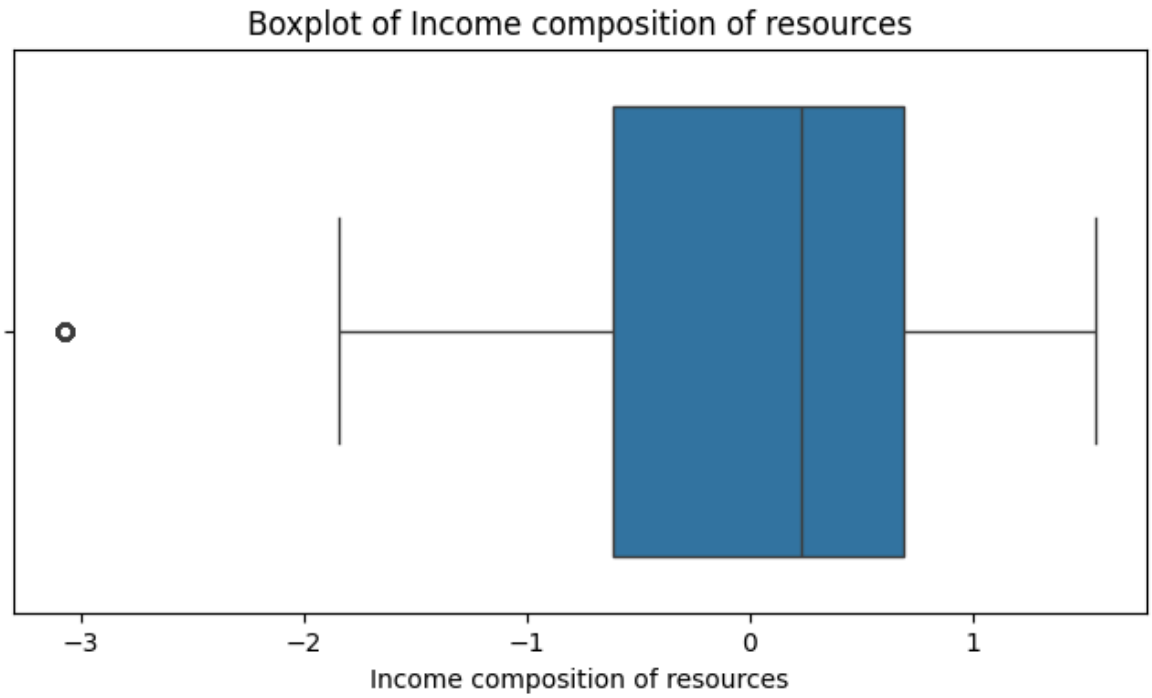

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seabor
n\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will
be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```



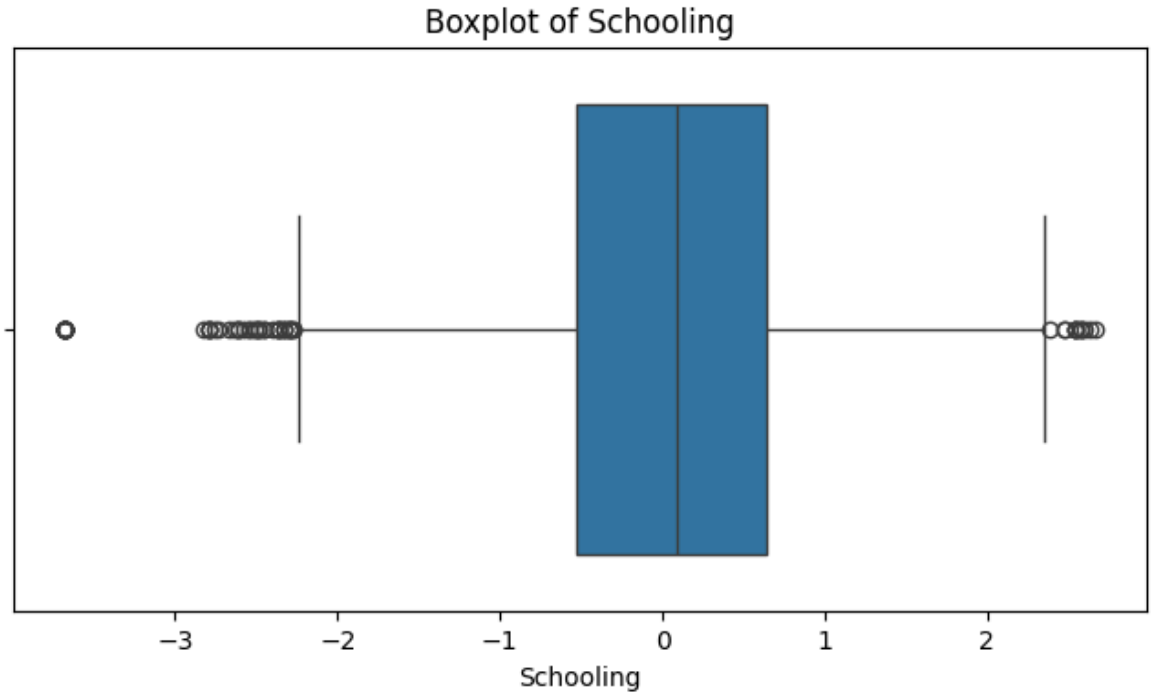

```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```




```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```




```
C:\Users\Gnanesh\AppData\Local\Programs\Python\Python312\Lib\site-packages\seaborn\categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be removed in a future version of pandas.
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```




In [ ]:
```
REMOVING OUTLIERS
```

In [14]:
```
for col in numeric_cols:
    Q1=df[col].quantile(0.25)
    Q3=df[col].quantile(0.75)
    IQR=Q3-Q1
    df=df[(df[col]>=Q1-1.5*IQR)&(df[col]<=Q3+1.5*IQR)]
print("\n Final dataset shape after removal of outliers:",df.shape)
```

```
Final dataset shape after removal of outliers: (623, 213)
```

In [15]:
```
df
```

Out[15]:

| | Year | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentage expenditure | Hepatitis B | |
|---|---|---|---|---|---|---|---|---|
| **16** | 1.621762 | 0.900898 | -0.731275 | -0.257017 | 0.013548 | -0.187805 | 0.694900 | -( |
| **17** | 1.404986 | 0.869344 | -1.263253 | -0.257017 | -0.009404 | -0.155718 | 0.651408 | -( |
| **18** | 1.188210 | 0.837790 | -0.650672 | -0.257017 | 0.054351 | -0.154648 | 0.694900 | -( |
| **19** | 0.971434 | 0.806236 | -0.634551 | -0.257017 | 0.151258 | -0.163922 | 0.694900 | -( |
| **20** | 0.754658 | 0.774682 | -0.618431 | -0.257017 | 0.209912 | -0.151536 | 0.694900 | -( |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **2858** | 1.621762 | 0.511731 | -0.062271 | -0.180685 | -0.201943 | -0.371433 | 0.173003 | -( |
| **2860** | 1.188210 | 0.480177 | -0.046150 | -0.180685 | 0.551638 | -0.371433 | -0.044454 | -( |
| **2861** | 0.971434 | 0.469659 | -0.030030 | -0.180685 | 0.549087 | -0.371433 | -0.087945 | -( |
| **2864** | 0.321106 | 0.459141 | 0.010272 | -0.180685 | 0.776054 | -0.371433 | 0.042529 | -( |
| **2869** | -0.762774 | 0.427587 | -0.030030 | -0.172204 | 0.765853 | -0.371433 | -0.044454 | -( |

623 rows × 213 columns

In [ ]: