In [105]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statistics
import math
from matplotlib.mlab import PCA as mlabPCA
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sqlalchemy import create_engine
import warnings
from scipy.stats.mstats import winsorize
from statsmodels.stats.weightstats import ttest_ind
import scipy
%matplotlib inline
from scipy import stats
from statsmodels.stats.weightstats import ttest_ind

postgres_user = 'dsbc_student'
postgres_pw = '7*.8G9QH21'
postgres_host = '142.93.121.174'
postgres_port = '5432'
postgres_db = 'lifeexpectancy'
table_name = 'lifeexpectancy'
```

In [2]:

```
life_df = pd.read_csv('lifeex.csv')
life_df.head()
```

Out[2]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
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

Dataset's content

The data-set related to life expectancy, health factors for 193 countries has been collected from WHO (World Health Organization) data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. In this project they have considered data from year 2000-2015 for the 193 countries for further analysis.

Challenge's goal

My goal in this challenge is to find the factors that affect the life expectancy. Specifically, I will need to find out which factors increase the expected life in the countries and which factors decrease it. The life expectancy variable is collected from the means of the expected factors that affects the life expectancy.

In [3]:
```
life_df.rename(columns = {' thinness  1-19 years': 'thinness 10-19 years'}, in
place = True)
print(life_df.columns)
```

```
Index(['Country', 'Year', 'Status', 'Life expectancy ', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B',
       'Measles ', ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditur
e',
       'Diphtheria ', ' HIV/AIDS', 'GDP', 'Population', 'thinness 10-19 year
s',
       ' thinness 5-9 years', 'Income composition of resources', 'Schoolin
g'],
      dtype='object')
```

I noticed a small error within the columns of the dataset. Before, one of the columns was named "thinness 1-19 years" for childrens of age 10-19. I just made a simple change to the name for correction. It only makes sense since we have a column named "thinness 5-9 years" for childrens of age 5-9, so it makes sense to have 10-19 years not 1-19 years.

My understanding

Our factor, life expectancy, have float values (Example: 65.0 in the first row). I believe this column's values are made from the mean of the necessary factors. The main question is, which factors were used to make the life expectancy variable.

Eliminating the necessary columns

Life expectancy is affected by many factors such as: socioeconomic status, including employment, income, education and economic wellbeing, the quality of the health system and the ability of people to access it, health behaviours such as tobacco and excessive alcohol consumption, poor nutrition and lack of exercise. With this in mind, I have to elimate the factors of which is irrelevant to the life expectancy. I wanted to take a look at the dataset once more to remove any columns that does not affect the topic variable.

In [4]:
```
life_df.head()
```

Out[4]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
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 [22]:
```
life_df2 = life_df.drop(['Population', 'under-five deaths ',
           'GDP', 'infant deaths', 'thinness 10-19 years', ' thinness 5-9 year
s', 'Adult Mortality'], axis=1)
life_df2.head()
```

Out[22]:

	Country	Year	Status	Life expectancy	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	Pol
0	Afghanistan	2015	Developing	65.0	0.01	71.279624	65.0	1154	19.1	6
1	Afghanistan	2014	Developing	59.9	0.01	73.523582	62.0	492	18.6	58
2	Afghanistan	2013	Developing	59.9	0.01	73.219243	64.0	430	18.1	62
3	Afghanistan	2012	Developing	59.5	0.01	78.184215	67.0	2787	17.6	67
4	Afghanistan	2011	Developing	59.2	0.01	7.097109	68.0	3013	17.2	68

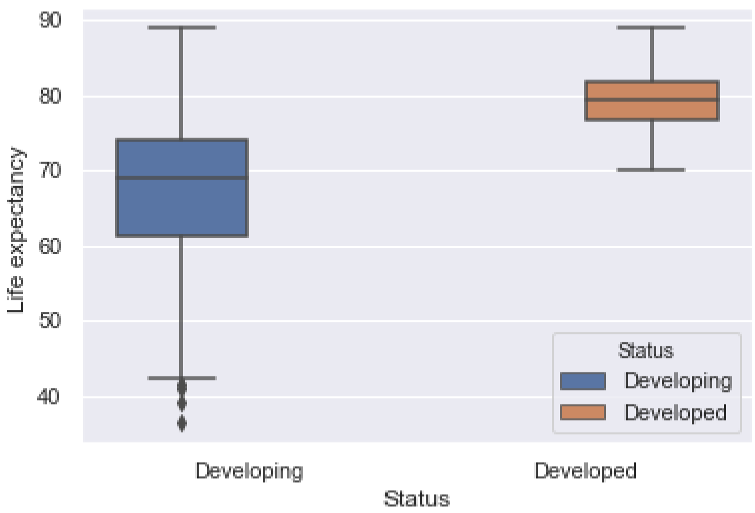
First thing's first, the categorical variable, "Status", which is the condition of the country. After looking through the "Status" column, I can see there are only 2 unique values. One is the "Developed" status and the other is "Developing" status. I expect the countries that are in development, or that are developing, may not have a higher life expectancy rate than the countries that are developed.

In [6]:

```
sns.set(style="darkgrid")

sns.boxplot(x="Status",y="Life expectancy ",
             hue="Status",
             data=life_df)

#recursive feature elimination
#featuring engineering
#T-testing or z-test

#modeling is next.
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x119ec96f0b8>
```



In the boxplot above, I can see both "Developing" and "Developed" status. As I expected, the countries that have the "Developing" status has an IQR between 61 and 74, whereas the countries with the "Developed" status, have an IQR between 78 and 82. This tells me the people in the countries that are well developed have a better life expectancy rate.

After taking another look at the dataset, I noticed there are a few variables in the dataset that surely have no affect on the life expectancy. Although the location does have an affect of the life expectancy, I shouldn't use the 'Country' variable because we have too many values within the variable to do an experiment. Also I excluded year because the time of year doesn't affect life expectancy directly. Next, is population which doesn't affect the topic variable directly. What matter is one's physical and mental health as well as economic wellbeing. Since deaths doesn't factor in life expectancy, I can exclude any variables that are related to deaths of people. I can exclude adult mortality because it is the probability of a 15 year old dying before age 60. I also excluded the thinness of children because I have the average of body mass as BMI.

Background details: Reason for the missing values

Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc.

In [7]:

```
#Check for missing values, make sure if there are any columns worth dropping.

life_df.isnull().mean()
```

Out[7]:

```
Country                            0.000000
Year                               0.000000
Status                             0.000000
Life expectancy                    0.003404
Adult Mortality                    0.003404
infant deaths                      0.000000
Alcohol                            0.066031
percentage expenditure             0.000000
Hepatitis B                        0.188223
Measles                            0.000000
 BMI                               0.011572
under-five deaths                  0.000000
Polio                              0.006467
Total expenditure                  0.076923
Diphtheria                         0.006467
 HIV/AIDS                          0.000000
GDP                                0.152485
Population                         0.221920
thinness 10-19 years               0.011572
 thinness 5-9 years                0.011572
Income composition of resources    0.056841
Schooling                          0.055480
dtype: float64
```

After checking all of the columns using simple coding, I can see we barely have any missing values in most of the columns. Most of the missing values are in the population, which is quite reasonable.

In [8]:

```
#fillna for all missing values

life_df["Life expectancy "].fillna(life_df["Life expectancy "].mean(), inplace=True)
life_df["Adult Mortality"].fillna(life_df["Adult Mortality"].mean(), inplace=True)
life_df["Alcohol"].fillna(life_df["Alcohol"].mean(), inplace=True)
life_df["Hepatitis B"].fillna(life_df["Hepatitis B"].mean(), inplace=True)
life_df[" BMI "].fillna(life_df[" BMI "].mean(), inplace=True)
life_df["Polio"].fillna(life_df["Polio"].mean(), inplace=True)
life_df["Total expenditure"].fillna(life_df["Total expenditure"].mean(), inplace=True)
life_df["Diphtheria "].fillna(life_df["Diphtheria "].mean(), inplace=True)
life_df["GDP"].fillna(life_df["GDP"].mean(), inplace=True)
life_df["Population"].fillna(life_df["Population"].mean(), inplace=True)

life_df["thinness 10-19 years"].fillna(life_df["thinness 10-19 years"].mean(), inplace=True)
life_df[" thinness 5-9 years"].fillna(life_df[" thinness 5-9 years"].mean(), inplace=True)
life_df["Income composition of resources"].fillna(life_df["Income composition of resources"].mean(), inplace=True)
life_df["Schooling"].fillna(life_df["Schooling"].mean(), inplace=True)

life_df.isnull().mean()
```

Out[8]:

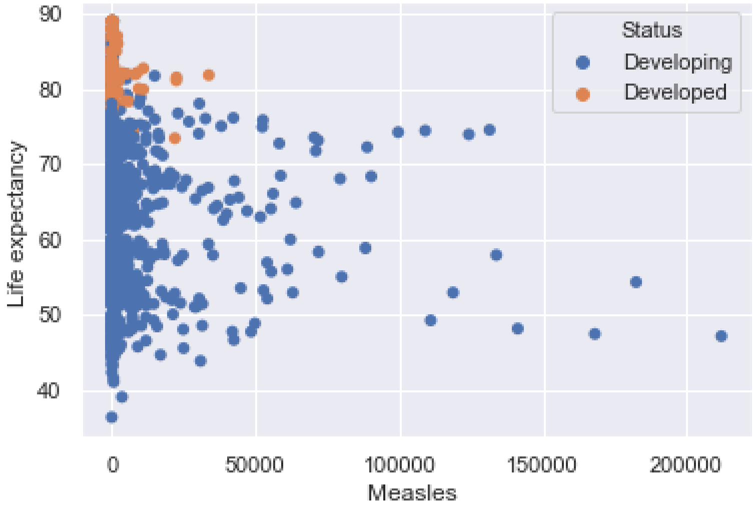
```
Country                            0.0
Year                               0.0
Status                             0.0
Life expectancy                    0.0
Adult Mortality                    0.0
infant deaths                      0.0
Alcohol                            0.0
percentage expenditure             0.0
Hepatitis B                        0.0
Measles                            0.0
 BMI                               0.0
under-five deaths                  0.0
Polio                              0.0
Total expenditure                  0.0
Diphtheria                         0.0
 HIV/AIDS                          0.0
GDP                                0.0
Population                         0.0
thinness 10-19 years               0.0
 thinness 5-9 years                0.0
Income composition of resources    0.0
Schooling                          0.0
dtype: float64
```

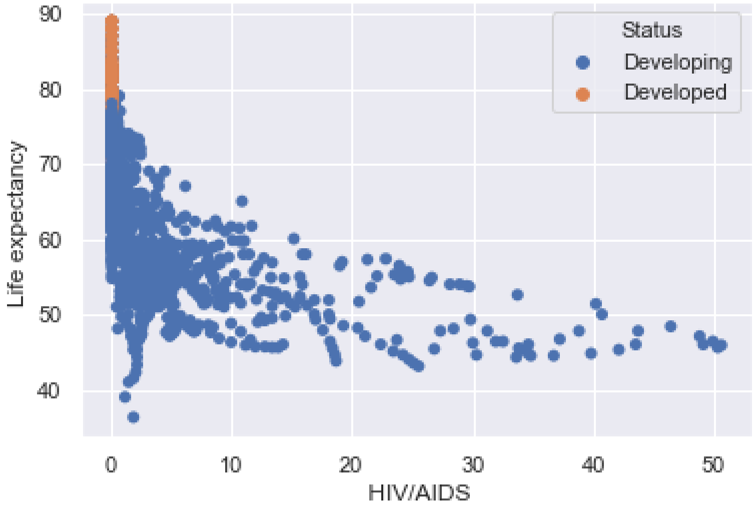
After a series of coding, I filled the missing values with another value. I figured this is the better option because there very little missing values so the data plots won't look much different with or without value fillups.

Do diseases affect the life expectancy rate?

We know people with diseases isn't a sign of good health. In fact, people sometimes pass away due to all kinds of diseases. I'm expecting the disease columns will have a huge impact of one's life expectancy with outliers as well.

In [26]:
```
sns.scatterplot(x="Measles ", y="Life expectancy ",
                hue="Status",
                sizes=(1, 8), linewidth=0,
                data=life_df2)
plt.show()

sns.scatterplot(x=" HIV/AIDS", y="Life expectancy ",
                hue="Status",
                sizes=(1, 8), linewidth=0,
                data=life_df2)
plt.show()
```





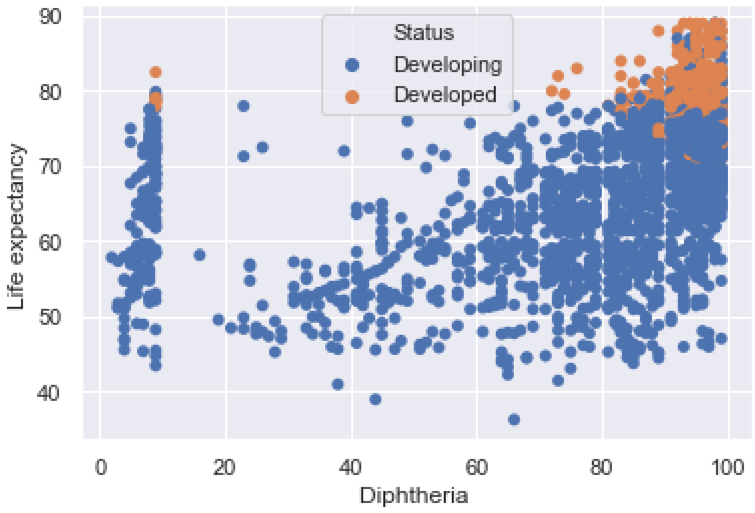
Looking at the disease scatterplots above, I can see the countries with the "Developed"(Orange values), status have little to none reported case of the measles as well as HIV/AIDS. There are a lot of developing countries and they have a lot of reported cases of the measles and HIV/AIDS, but the plots above confirm the developed countries have a higher life expectancy rate because they have little to none reported cases of diseases.

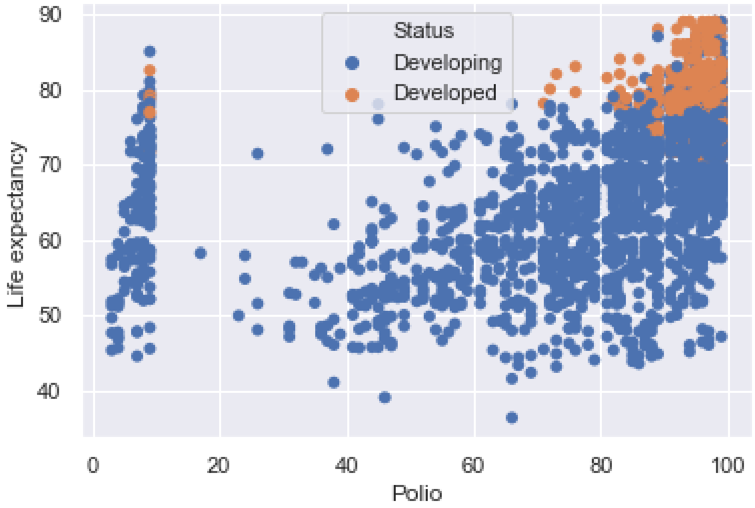
Does the disease immunization coverages affect the life expectancy as well?

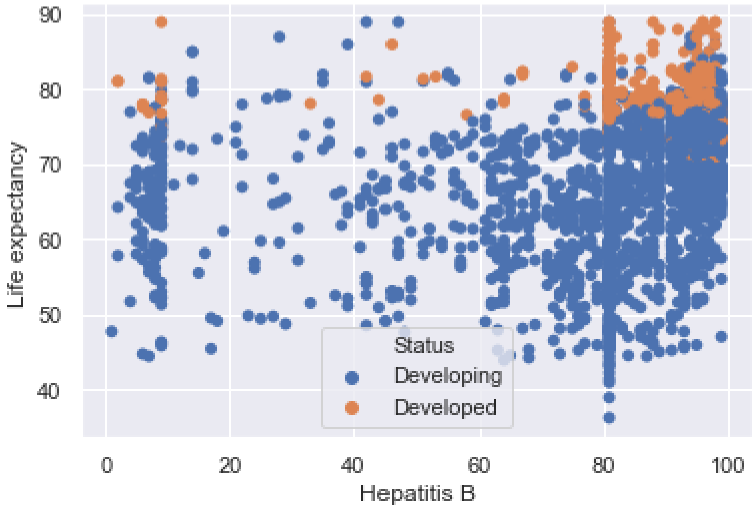
When talking about the immunization coverages, we are taking a look at the Diphtheria, Polio and Hepatitis B continuous variable columns, with 0% means no coverage, and 100% being fully covered. I expect the "Developed" countries to outbeat the "Developing countries in terms of life expectancy as well.

In [25]:

```
sns.scatterplot(x="Diphtheria ", y="Life expectancy ",
                hue="Status",
                sizes=(1, 8), linewidth=0,
                data=life_df2)
plt.show()

sns.scatterplot(x="Polio", y="Life expectancy ",
                hue="Status",
                sizes=(1, 8), linewidth=0,
                data=life_df2)
plt.show()

sns.scatterplot(x="Hepatitis B", y="Life expectancy ",
                hue="Status",
                sizes=(1, 8), linewidth=0,
                data=life_df2)
plt.show()
```







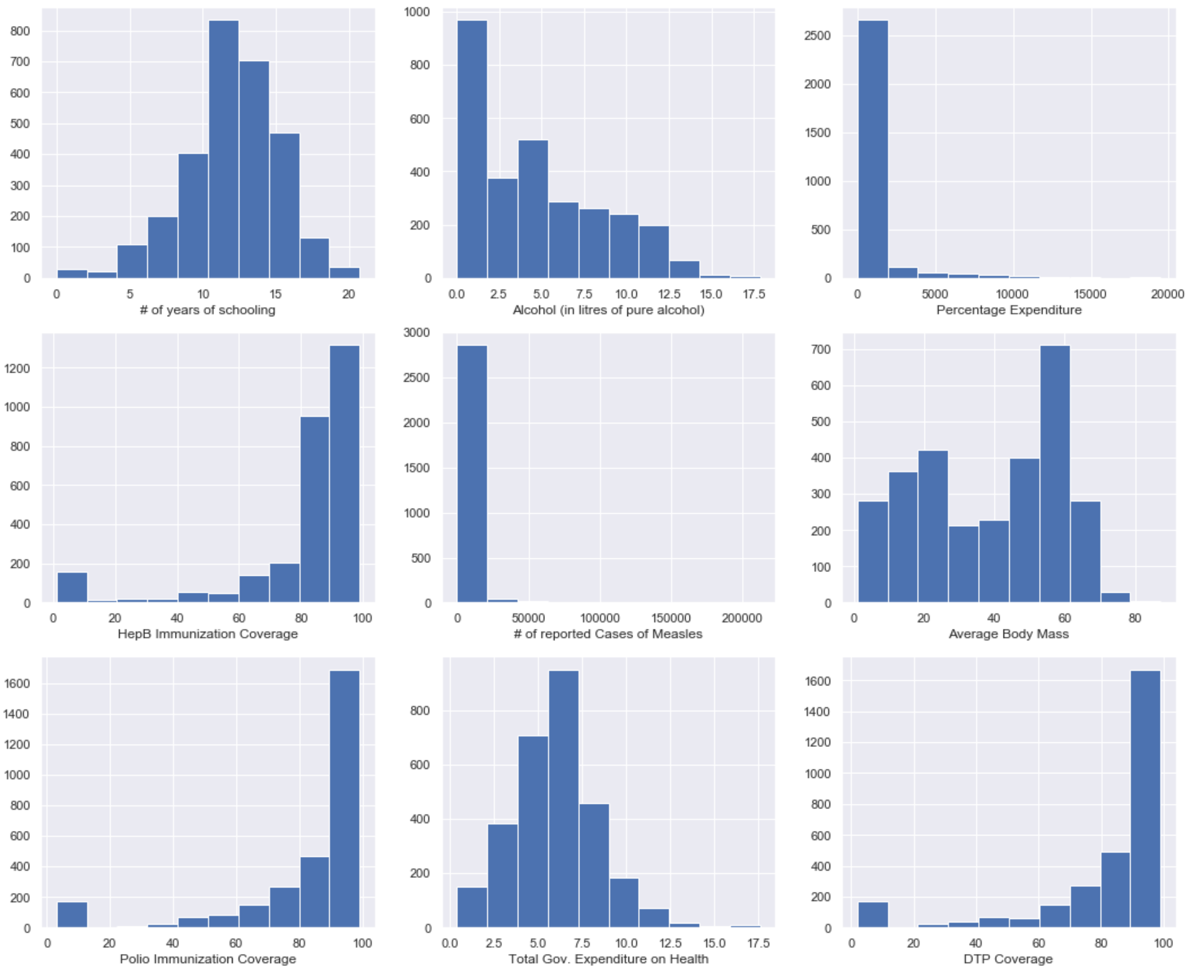
Looking at all three scatterplots above, the immunization coverage for all three diseases (Diphtheria, Polio, Hepatitis B), it shows us the a lot of people that are covered to have a higher life expectancy rate than the ones that do not. There are some people that aren't covered and have a high rate as well but I expect these kinds of outliers since all people aren't the same. Some can live longer with and without the immunization coverage.

Dealing with the outliers

Now I must make sure if there are any outliers. First, I'm expecting there to be outliers since people live longer than others. Some live a lot long than others, some live a lot shorter than others. I checked the topic's data, the Life expectancy column, and had to figure out what is giving the life expectancy variable the outliers? Then I'd have a better idea about which other variable is affecting the topic variable (life expectancy).

In [24]:

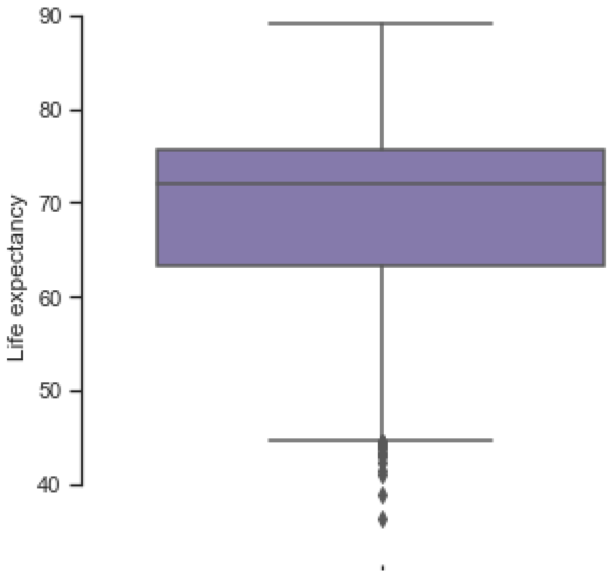
```
sns.set(style="darkgrid")

plt.figure(figsize=(18,15))

plt.subplot(3,3,1)
plt.hist(life_df2["Schooling"])
plt.xlabel("# of years of schooling")

plt.subplot(3,3,2)
plt.hist(life_df2["Alcohol"])
plt.xlabel("Alcohol (in litres of pure alcohol)")

plt.subplot(3,3,3)
plt.hist(life_df2["percentage expenditure"])
plt.xlabel("Percentage Expenditure")

plt.subplot(3,3,4)
plt.hist(life_df2["Hepatitis B"])
plt.xlabel("HepB Immunization Coverage")

plt.subplot(3,3,5)
plt.hist(life_df2["Measles "])
plt.xlabel("# of reported Cases of Measles")

plt.subplot(3,3,6)
plt.hist(life_df2[" BMI "])
plt.xlabel("Average Body Mass")

plt.subplot(3,3,7)
plt.hist(life_df2["Polio"])
plt.xlabel("Polio Immunization Coverage")

plt.subplot(3,3,8)
plt.hist(life_df2["Total expenditure"])
plt.xlabel("Total Gov. Expenditure on Health")

plt.subplot(3,3,9)
plt.hist(life_df2["Diphtheria "])
plt.xlabel("DTP Coverage")

plt.show()
```



Judging from all the other outliers, I've assumed the Life expectancy would have outliers as well. I also noticed some of the histograms of the original variables demonstrate skewed distributions. It may not be necessary to get rid of the outliers in this case.

In [12]:

```
sns.set(style="ticks")

plt.figure(figsize=(5,5))

sns.boxplot(y="Life expectancy ",
            palette=["m"],
            data=life_df)
sns.despine(offset=10, trim=True)
```



As expected, the life expectancy's boxplot shows the variable has outliers.

In [13]:

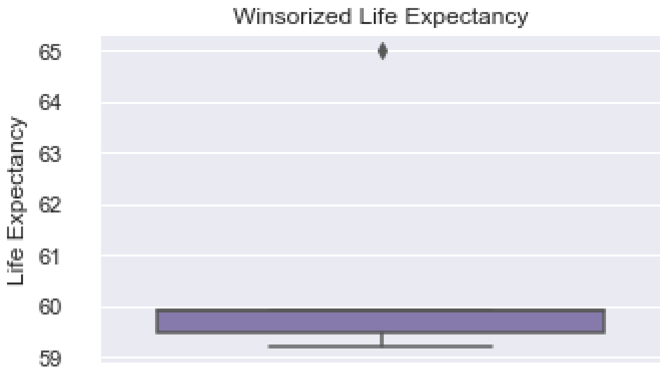
```
winsorized_life_expectancy = winsorize(life_df2["Life expectancy "], (0, 0))

winsorized_life_expectancy
```

Out[13]:

```
masked_array(data=[65. , 59.9, 59.9, 59.5, 59.2],
             mask=False,
       fill_value=1e+20)
```

In the "dealing with outliers" phase, I had to think if I should remove the outliers in the factors that gave the life expectancy's outliers, or go for the main topic and remove the life expectancy outliers. After careful thought, I found it was better to deal with the life expectancy's outliers by using the winsorizing method. Winsorizing the life expectancy's boxplot will remove all of the data points that are not even close than the others.

```
In [14]: sns.set(style="darkgrid")

         plt.figure(figsize=(5,3))

         sns.boxplot(y= winsorized_life_expectancy,
                     palette=["m"],
                     data=life_df2)
         sns.despine(offset=10, trim=True)
         plt.ylabel("Life Expectancy")
         plt.title("Winsorized Life Expectancy")

         plt.show()
```



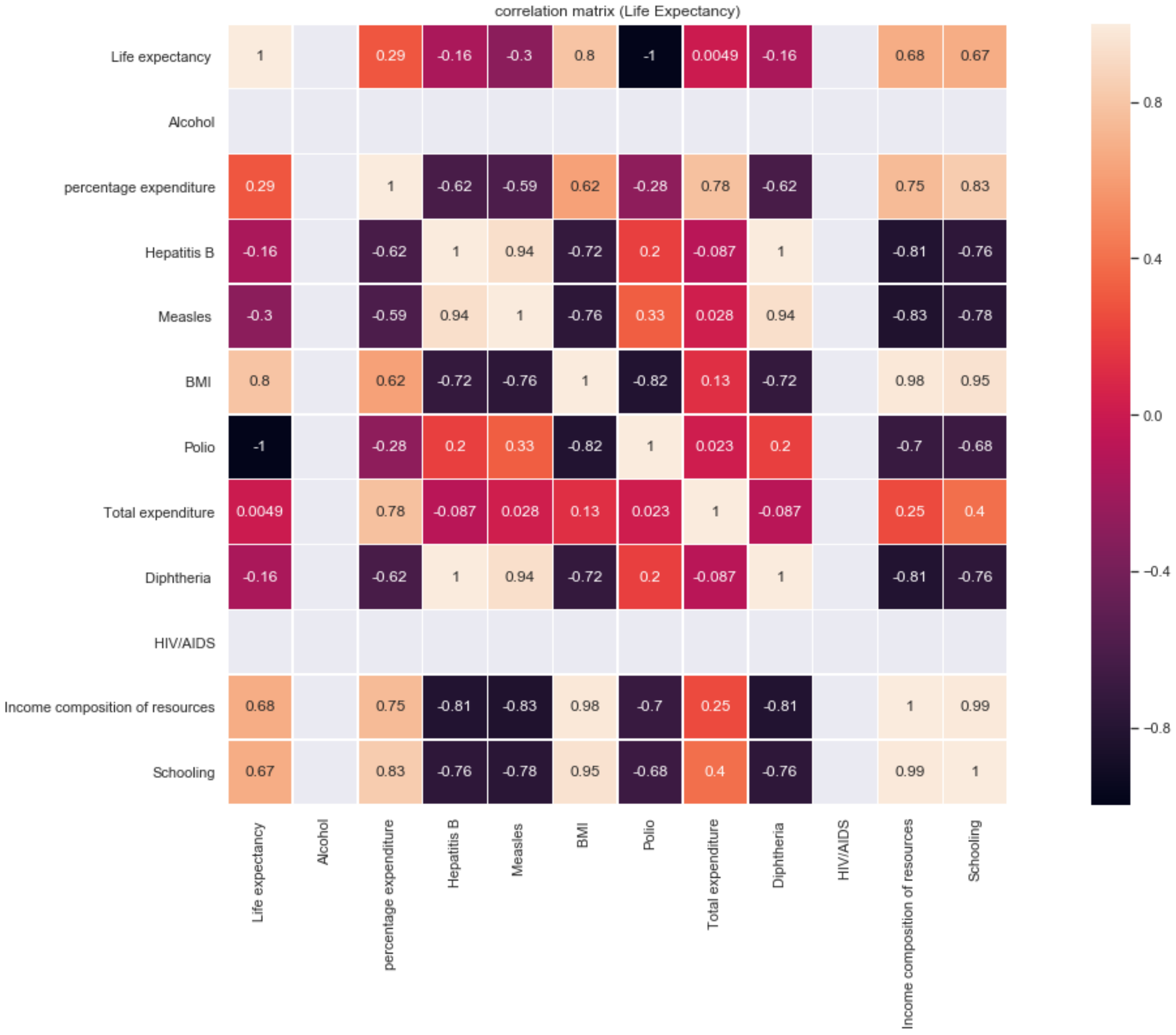
After creating a boxplot for our dataset's Life expectancy column, I've revealed there are a lot of outliers. I used the winsorization method so drop most of the outliers and it returned one outlier of 65.0 which is closer to the 59.9s' (the IQR). I was expecting this outcome because there will be outliers no matter what since people live longer or shorter than others.

Correlation with the Topic Variable

Since I am looking at the Life expectancy variable, I will check to see which variables are most correlated with the life expectancy variable.

In [15]:
```
corrmat_life_df2 = life_df2.corr()

plt.subplots(figsize=(23, 11))

sns.heatmap(corrmat_life_df2, square=True, annot=True, linewidths=.5)
plt.title("correlation matrix (Life Expectancy)")
```

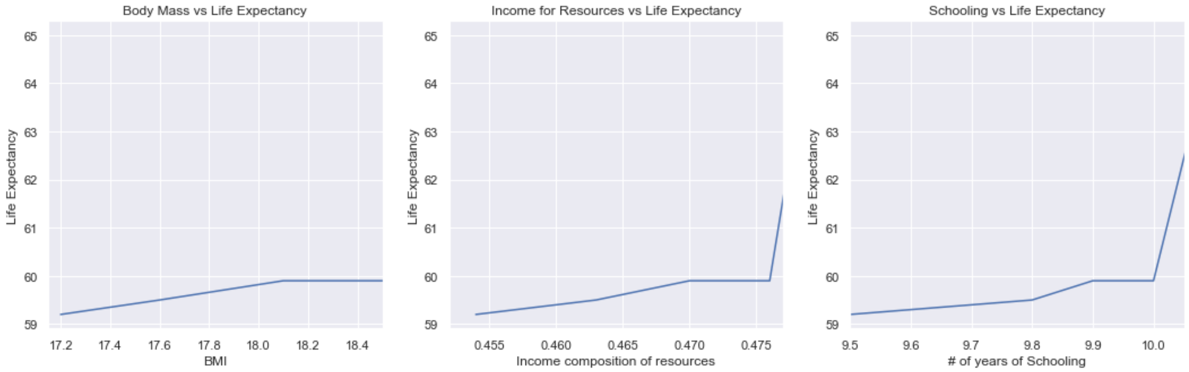
Out[15]:
```
Text(0.5, 1.0, 'correlation matrix (Life Expectancy)')
```



After creating a heatmap for correlation comparison, I can see BMI, Income composition of resources and schooling are all mostly correlated with the life expectancy variable than the others. I'm expecting those three variables will have an impact on the life expectancy.

In [16]:

```
sns.set(style="darkgrid")
plt.figure(figsize=(18,5))

plt.subplot(1,3,1)
plt.plot(life_df2[" BMI "], life_df2["Life expectancy "])
plt.title("Body Mass vs Life Expectancy")
plt.xlabel("BMI")
plt.ylabel("Life Expectancy")
plt.xlim([17.15, 18.5])

plt.subplot(1,3,2)
plt.plot(life_df2["Income composition of resources"], life_df2["Life expectanc
y "])
plt.title("Income for Resources vs Life Expectancy")
plt.xlabel("Income composition of resources")
plt.ylabel("Life Expectancy")
plt.xlim([0.452, 0.477])

plt.subplot(1,3,3)
plt.plot(life_df2["Schooling"], life_df2["Life expectancy "])
plt.title("Schooling vs Life Expectancy")
plt.xlabel("# of years of Schooling")
plt.ylabel("Life Expectancy")
plt.xlim([9.5, 10.05])

plt.show()
```



After creating a lineplot for BMI, Income composition of resources, Schooling and the winsorized life expectancy variable, I concluded the higher the BMI, Income composition of resource and Schooling goes, the higher the life expectancy goes as well.

Summary

The important factors that affects the life expectancy are a healthy body mass, income for resources and education. With education, you can learn how to solve lots of problems, not to mention it will keep you busy and out of trouble. If you are well educated, you may feel good about yourself. You will know all of the healthy food to put in your body and know what kind of exercise routine to take on.

Money is also an important thing to have. If you are financially stable, it can affect your mood in a good way. You may be less stressed if you always have the necessary resources, bought by money you need to live and support others. If an unfortunate accident occurs, you will have less worries on the cost of the health recovery. When talking about national income, leaders can use income to develop health programs for people to increase their life expectancy as well. There are many reasons why income for resources is an important factor for one's life expectancy.

Having an unhealthy body mass can lead to a lot of problems. Since humans are so fragile, and if the body isn't well taken care of, it can lead to heart problems, blood problems, emotional problems, and many more health problems. There's only so much mass the average body can carry. Having very little body mass can be a problem and having a lot of body mass can be a problem as well.

That is why all three of these factors, including diseases and their coverages, are the most important factors for one's life expectancy.