EDA and Data Wrangling (HOA 11.1)

In [213]:
```
import pandas as pd

life_expectancy_df = pd.read_csv('/content/data/Life Expectancy Data.csv')
life_expectancy_df.head()
```

Out[213]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphth
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	

5 rows × 22 columns

In [214]:
```
life_expectancy_df.shape
```

Out[214]:
```
(2938, 22)
```

In [215]:
```
life_expectancy_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 [216]:
```
life_expectancy_df.describe()
```

Out[216]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	unc
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.

Making a copy of the original dataset.

In [217]:
```
life_expectancy_df_raw = life_expectancy_df.copy()
```

Data Cleaning

Changing the column names by removing the white spaces.

In [218]:
```
life_expectancy_df.columns
```

Out[218]:
```
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')
```

In [240]:
```
life_expectancy_df.rename(columns={
    'Country': 'country',
    'Year': 'year',
    'Status': 'status',
    'Life expectancy ': ' life_expectancy',
    'Adult Mortality': ' adult_mortality',
    'infant deaths': 'infant_deaths',
    'Alcohol': 'alcohol',
    'percentage expendeture': 'percentage_expendeture',
    'Hepatitis B': 'hepatitis_b',
    'Measles ': 'measles',
    ' BMI ': 'BMI',
    'under-five deaths ': 'under_five_deaths',
    'Polio': 'polio',
    'Total expenditure': 'total_expenditure',
    'Diphtheria ': 'diphtheria',
    ' HIV/AIDS': 'HIV_AIDS',
    'Population': 'population',
    ' thinness  1-19 years': 'thinness_1-19_years',
    ' thinness 5-9 years': 'thinness_5-9_years',
    'Income composition of resources': 'income_composition_of_resources',
    'Schooling': 'schooling'
}, inplace=True)

life_expectancy_df.columns
```

Out[240]:
```
Index(['country', 'year', 'status', ' life_expectancy', ' adult_mortality',
       'infant_deaths', 'alcohol', 'percentage expenditure', 'measles',
       'under_five_deaths', 'polio', 'total_expenditure', 'diphtheria',
       'HIV_AIDS', 'GDP', 'thinness_1-19_years', 'thinness_5-9_years',
       'income_composition_of_resources', 'schooling'],
      dtype='object')
```

Fixing the countries issue.

In [220]:
```
min_row_per_country = 16

# number of rows per country
rows_per_country = life_expectancy_df['country'].value_counts()

# filtering countries that didn't meet the soecified min
countries_to_keep = rows_per_country[rows_per_country >= min_row_per_country].index

# filtering the dataframe
life_expectancy_df = life_expectancy_df[life_expectancy_df['country'].isin(countries_to_keep)]

life_expectancy_df['country'].isnull().sum()
```

Out[220]:
```
0
```

Handling missing values.

In [221]:
```
life_expectancy_df.isnull().sum()
```

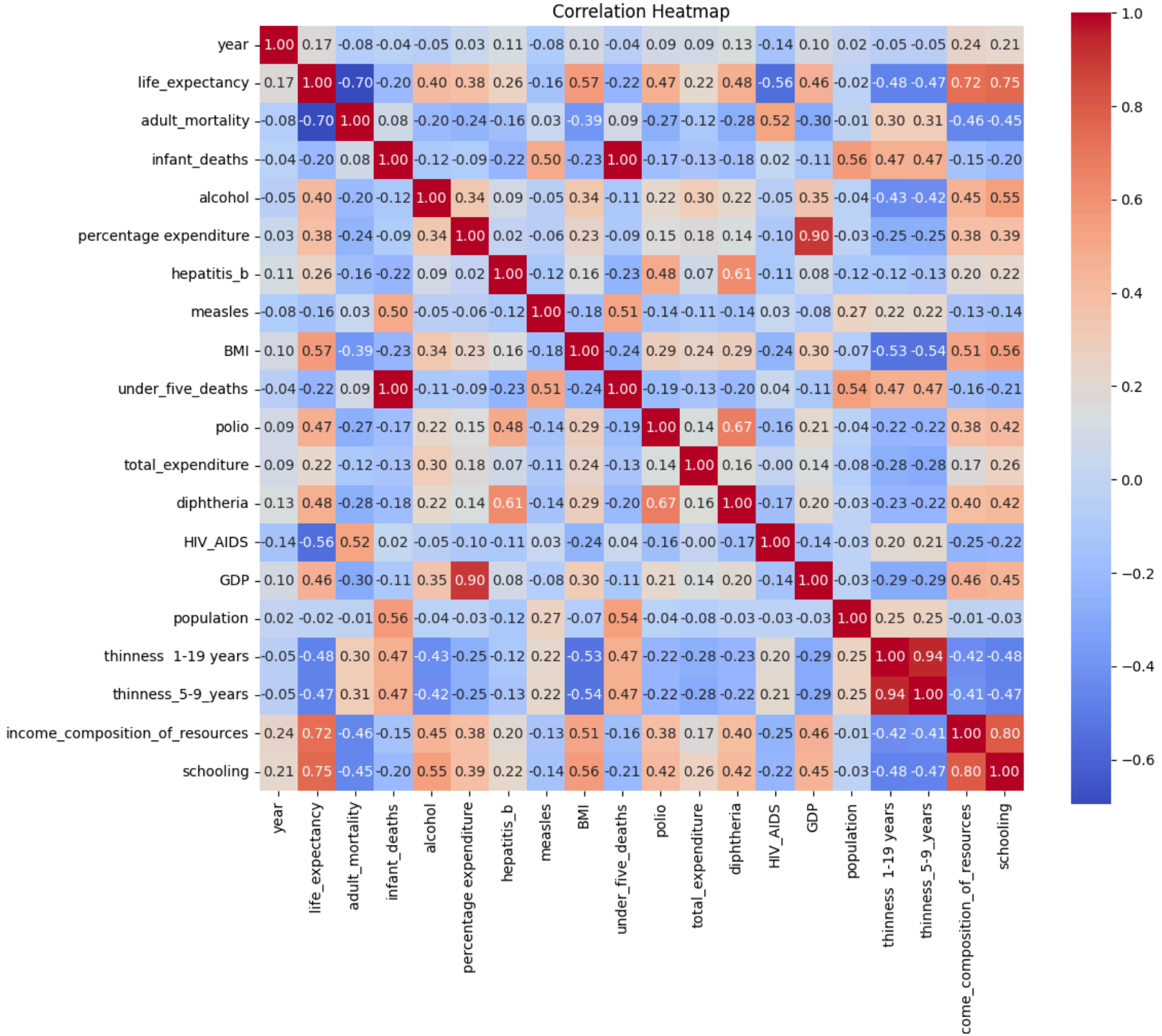
Out[221]:
```
country                             0
year                                0
status                              0
 life_expectancy                    0
 adult_mortality                    0
infant_deaths                       0
alcohol                           193
percentage expenditure              0
hepatitis_b                       553
measles                             0
BMI                                32
under_five_deaths                   0
polio                              19
total_expenditure                 226
diphtheria                         19
HIV_AIDS                            0
GDP                               443
population                        644
 thinness  1-19 years              32
thinness_5-9_years                 32
income_composition_of_resources   160
schooling                         160
dtype: int64
```

In [222]:

```
import seaborn as sns
import matplotlib.pyplot as plt

num_cols = life_expectancy_df.select_dtypes(include=['float64', 'int64']).columns
correlation_matrix = life_expectancy_df[num_cols].corr()

plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm', square=True)
plt.title('Correlation Heatmap')
plt.show()
```


Correlation Heatmap
year 1.00 0.17 -0.08 -0.04 -0.05 0.03 0.11 -0.08 0.10 -0.04 0.09 0.09 0.13 -0.14 0.10 0.02 -0.05 -0.05 0.24 0.21
life_expectancy 0.17 1.00 -0.70 -0.20 0.40 0.38 0.26 -0.16 0.57 -0.22 0.47 0.22 0.48 -0.56 0.46 -0.02 -0.48 -0.47 0.72 0.75
adult_mortality -0.08 -0.70 1.00 0.08 -0.20 -0.24 -0.16 0.03 -0.39 0.09 -0.27 -0.12 -0.28 0.52 -0.30 -0.01 0.30 0.31 -0.46 -0.45
infant_deaths -0.04 -0.20 0.08 1.00 -0.12 -0.09 -0.22 0.50 -0.23 1.00 -0.17 -0.13 -0.18 0.02 -0.11 0.56 0.47 0.47 -0.15 -0.20
alcohol -0.05 0.40 -0.20 -0.12 1.00 0.34 0.09 -0.05 0.34 -0.11 0.22 0.30 0.22 -0.05 0.35 -0.04 -0.43 -0.42 0.45 0.55
percentage expenditure 0.03 0.38 -0.24 -0.09 0.34 1.00 0.02 -0.06 0.23 -0.09 0.15 0.18 0.14 -0.10 0.90 -0.03 -0.25 -0.25 0.38 0.39
hepatitis_b 0.11 0.26 -0.16 -0.22 0.09 0.02 1.00 -0.12 0.16 -0.23 0.48 0.07 0.61 -0.11 0.08 -0.12 -0.12 -0.13 0.20 0.22
measles -0.08 -0.16 0.03 0.50 -0.05 -0.06 -0.12 1.00 -0.18 0.51 -0.14 -0.11 -0.14 0.03 -0.08 0.27 0.22 0.22 -0.13 -0.14
BMI 0.10 0.57 -0.39 -0.23 0.34 0.23 0.16 -0.18 1.00 -0.24 0.29 0.24 0.29 -0.24 0.30 -0.07 -0.53 -0.54 0.51 0.56
under_five_deaths -0.04 -0.22 0.09 1.00 -0.11 -0.09 -0.23 0.51 -0.24 1.00 -0.19 -0.13 -0.20 0.04 -0.11 0.54 0.47 0.47 -0.16 -0.21
polio 0.09 0.47 -0.27 -0.17 0.22 0.15 0.48 -0.14 0.29 -0.19 1.00 0.14 0.67 -0.16 0.21 -0.04 -0.22 -0.22 0.38 0.42
total_expenditure 0.09 0.22 -0.12 -0.13 0.30 0.18 0.07 -0.11 0.24 -0.13 0.14 1.00 0.16 -0.00 0.14 -0.08 -0.28 -0.28 0.17 0.26
diphtheria 0.13 0.48 -0.28 -0.18 0.22 0.14 0.61 -0.14 0.29 -0.20 0.67 0.16 1.00 -0.17 0.20 -0.03 -0.23 -0.22 0.40 0.42
HIV_AIDS -0.14 -0.56 0.52 0.02 -0.05 -0.10 -0.11 0.03 -0.24 0.04 -0.16 -0.00 -0.17 1.00 -0.14 -0.03 0.20 0.21 -0.25 -0.22
GDP 0.10 0.46 -0.30 -0.11 0.35 0.90 0.08 -0.08 0.30 -0.11 0.21 0.14 0.20 -0.14 1.00 -0.03 -0.29 -0.29 0.46 0.45
population 0.02 -0.02 -0.01 0.56 -0.04 -0.03 -0.12 0.27 -0.07 0.54 -0.04 -0.08 -0.03 -0.03 -0.03 1.00 0.25 0.25 -0.01 -0.03
thinness 1-19 years -0.05 -0.48 0.30 0.47 -0.43 -0.25 -0.12 0.22 -0.53 0.47 -0.22 -0.28 -0.23 0.20 -0.29 0.25 1.00 0.94 -0.42 -0.48
thinness_5-9_years -0.05 -0.47 0.31 0.47 -0.42 -0.25 -0.13 0.22 -0.54 0.47 -0.22 -0.28 -0.22 0.21 -0.29 0.25 0.94 1.00 -0.41 -0.47
income_composition_of_resources 0.24 0.72 -0.46 -0.15 0.45 0.38 0.20 -0.13 0.51 -0.16 0.38 0.17 0.40 -0.25 0.46 -0.01 -0.42 -0.41 1.00 0.80
schooling 0.21 0.75 -0.45 -0.20 0.55 0.39 0.22 -0.14 0.56 -0.21 0.42 0.26 0.42 -0.22 0.45 -0.03 -0.48 -0.47 0.80 1.00
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
come_composition_of_resources
schooling
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Replacing alcohol and total_expendeture with missing values with the mean because there is a strong positive correleation between the alcohol and total_expenditure despite having missing values.

In [223]:
```
life_expectancy_df[pd.isnull(life_expectancy_df['alcohol'])].head()
```

Out[223]:

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	hepatitis_b	measles	...	polio	tot
32	Algeria	2015	Developing	75.6	19.0	21	NaN	0.0	95.0	63	...	95.0	
48	Angola	2015	Developing	52.4	335.0	66	NaN	0.0	64.0	118	...	7.0	
64	Antigua and Barbuda	2015	Developing	76.4	13.0	0	NaN	0.0	99.0	0	...	86.0	
80	Argentina	2015	Developing	76.3	116.0	8	NaN	0.0	94.0	0	...	93.0	
96	Armenia	2015	Developing	74.8	118.0	1	NaN	0.0	94.0	33	...	96.0	

5 rows × 22 columns

In [224]:
```
# mean of alcohol and total_expenditure columns
alcohol_mean = life_expectancy_df['alcohol'].mean()
total_ex_mean = life_expectancy_df['total_expenditure'].mean()

# replacing nan values with their mean
life_expectancy_df['alcohol'].fillna(alcohol_mean, inplace=True)
life_expectancy_df['total_expenditure'].fillna(total_ex_mean, inplace=True)

life_expectancy_df.head()
```

Out[224]:

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	hepatitis_b	measles	...	polio	tc
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	

5 rows × 22 columns

In [225]:
```
life_expectancy_df[['alcohol', 'total_expenditure']].isnull().sum()
```

Out[225]:
```
alcohol              0
total_expenditure    0
dtype: int64
```

There is a low correlation between the hepatitis_b and the life_expectancy column which makes them insignificant. Given the large amount of null values hepatitis_b contains, I think it should be dropped.

In [226]:
```
life_expectancy_df = life_expectancy_df.drop('hepatitis_b', axis=1)
```

Missing values in the polio, GDP, and diphtheria column will be replaced with the mean values based on countries.

In [227]:
```
# gpd value mean for each country
mean_gdp_val = life_expectancy_df.groupby('country')['GDP'].mean()

# replacing mean values with the mean per country
life_expectancy_df['GDP'] = life_expectancy_df['GDP'].fillna(life_expectancy_df.groupby('country')['GDP'].tra

# Getting the remaining NaN value counts
life_expectancy_df['GDP'].isnull().sum()
```

Out[227]:
```
400
```

In [228]:
```
# dropping the remaining rows with nan values
life_expectancy_df = life_expectancy_df.dropna(subset=['GDP'])

life_expectancy_df['GDP'].isnull().sum()
```

Out[228]:
```
0
```

In [229]:
```
# gpd value mean for each country
mean_gdp_val = life_expectancy_df.groupby('country')['polio'].mean()

# replacing mean values with the mean per country
life_expectancy_df['polio'] = life_expectancy_df['polio'].fillna(life_expectancy_df.groupby('country')['polio

# Getting the remaining NaN value counts
life_expectancy_df['polio'].isnull().sum()
```

Out[229]:
```
0
```

In [230]:
```
life_expectancy_df.columns
```

Out[230]:
```
Index(['country', 'year', 'status', ' life_expectancy', ' adult_mortality',
       'infant_deaths', 'alcohol', 'percentage expenditure', '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 [232]:
```
# gpd value mean for each country
mean_gdp_val = life_expectancy_df.groupby('country')['diphtheria'].mean()

# replacing mean values with the mean per country
life_expectancy_df['diphtheria'] = life_expectancy_df['diphtheria'].fillna(life_expectancy_df.groupby('countr

# Getting the remaining NaN value counts
life_expectancy_df['diphtheria'].isnull().sum()
```

Out[232]:
```
0
```

Much like the hepatitis_b column, population and BMI column are quite insignificant which leads me to dropping it.

In [233]:
```
life_expectancy_df = life_expectancy_df.drop('population', axis=1)
```

Out[233]:

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	measles	BMI	under_five_deaths
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	1154	19.1	83
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	492	18.6	86
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	430	18.1	89
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	2787	17.6	93
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	3013	17.2	97

In [ ]:
```
life_expectancy_df = life_expectancy_df.drop('BMI', axis=1)
```

All null values in the thinness_1-19_years column is in the country South Sudan and will be dropped.

In [242]:
```
life_expectancy_df[life_expectancy_df['thinness_1-19_years'].isnull()].head()
```

Out[242]:

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	measles	under_five_deaths	poli
2409	South Sudan	2015	Developing	57.3	332.0	26	4.614856	0.000000	878	39	41
2410	South Sudan	2014	Developing	56.6	343.0	26	4.614856	46.074469	441	39	44
2411	South Sudan	2013	Developing	56.4	345.0	26	4.614856	47.444530	525	40	5
2412	South Sudan	2012	Developing	56.0	347.0	26	4.614856	38.338232	1952	40	64
2413	South Sudan	2011	Developing	55.4	355.0	27	4.614856	0.000000	1256	41	66

In [248]:
```
life_expectancy_df = life_expectancy_df[life_expectancy_df['country'] != 'South Sudan']
```

Dropping remaining null values.

In [250]:
```
life_expectancy_df.dropna(subset=['income_composition_of_resources', 'schooling'], inplace=True)
```

In [251]:
```
life_expectancy_df.isnull().sum()
```

Out[251]:
```
country                          0
year                             0
status                           0
 life_expectancy                 0
 adult_mortality                 0
infant_deaths                    0
alcohol                          0
percentage expenditure           0
measles                          0
under_five_deaths                0
polio                            0
total_expenditure                0
diphtheria                       0
HIV_AIDS                         0
GDP                              0
thinness_1-19_years              0
thinness_5-9_years               0
income_composition_of_resources  0
schooling                        0
dtype: int64
```

In [252]:
```
# exporting to csv file
life_expectancy_df.to_csv('data/life_expectancy_clean.csv')
```

EDA

In [254]:
```
life_expectancy_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2480 entries, 0 to 2937
Data columns (total 19 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   country                          2480 non-null   object 
 1   year                             2480 non-null   int64  
 2   status                           2480 non-null   object 
 3    life_expectancy                 2480 non-null   float64
 4    adult_mortality                 2480 non-null   float64
 5   infant_deaths                    2480 non-null   int64  
 6   alcohol                          2480 non-null   float64
 7   percentage expenditure           2480 non-null   float64
 8   measles                          2480 non-null   int64  
 9   under_five_deaths                2480 non-null   int64  
 10  polio                            2480 non-null   float64
 11  total_expenditure                2480 non-null   float64
 12  diphtheria                       2480 non-null   float64
 13  HIV_AIDS                         2480 non-null   float64
 14  GDP                              2480 non-null   float64
 15  thinness_1-19_years              2480 non-null   float64
 16  thinness_5-9_years               2480 non-null   float64
 17  income_composition_of_resources  2480 non-null   float64
 18  schooling                        2480 non-null   float64
dtypes: float64(13), int64(4), object(2)
memory usage: 387.5+ KB
```

In [257]:
```
life_expectancy_df.describe()
```

Out[257]:

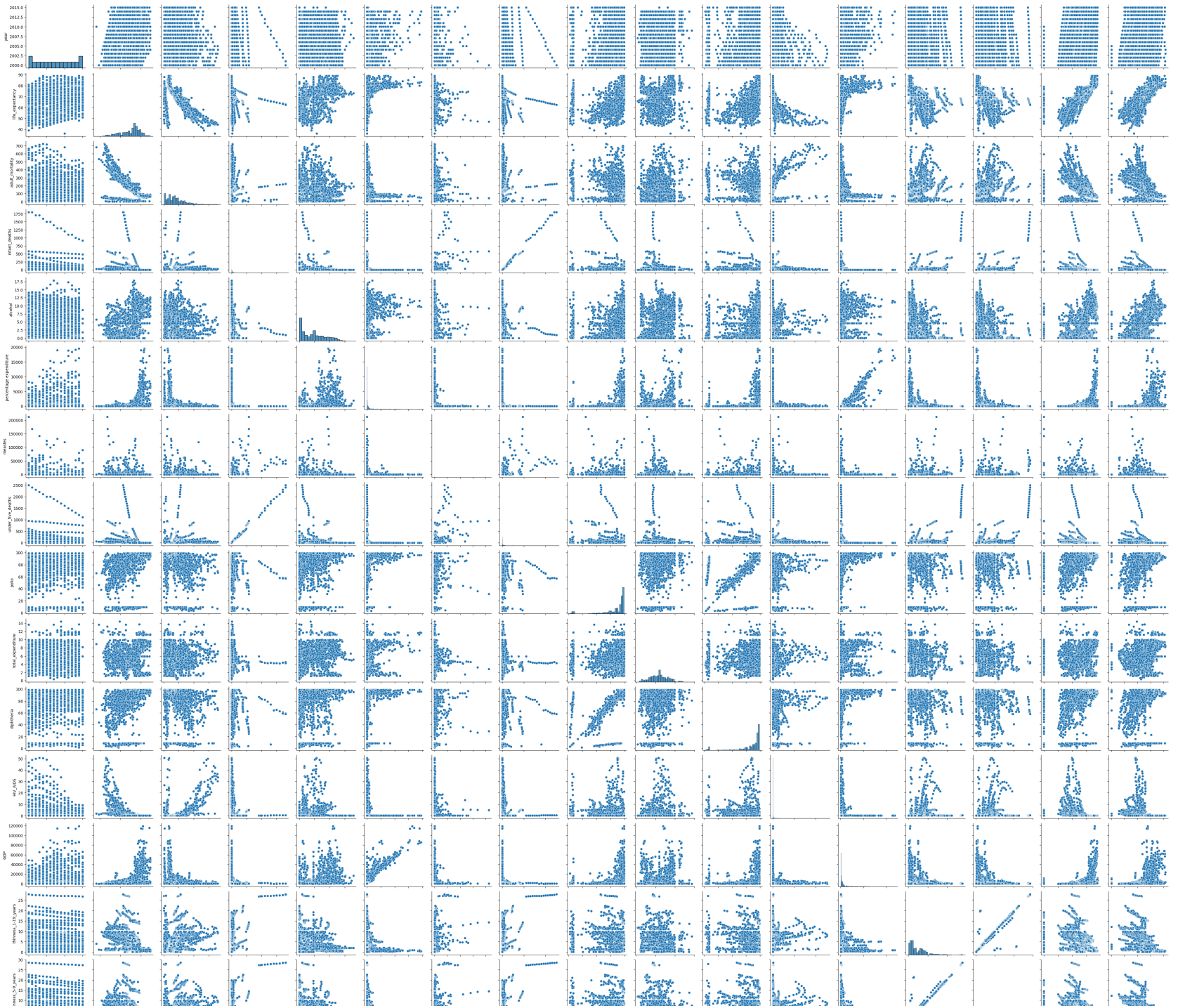
	year	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage expenditure	measles	under_five_deaths	
count	2480.000000	2480.000000	2480.000000	2480.000000	2480.000000	2480.000000	2480.000000	2480.000000	2480.0
mean	2007.500000	69.472863	160.880645	30.921371	4.580352	873.373144	2338.379839	42.884274	82.5
std	4.610702	9.598148	125.819455	126.778440	3.905404	2136.417700	11106.210149	172.351941	23.2
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	0.000000	0.000000	3.0
25%	2003.750000	63.500000	71.000000	0.000000	0.990000	23.654806	0.000000	0.000000	78.0
50%	2007.500000	72.300000	137.000000	2.000000	4.250000	119.901020	15.000000	3.000000	93.0
75%	2011.250000	76.000000	223.000000	18.000000	7.330000	576.437462	334.000000	23.000000	97.0
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	212183.000000	2500.000000	99.0

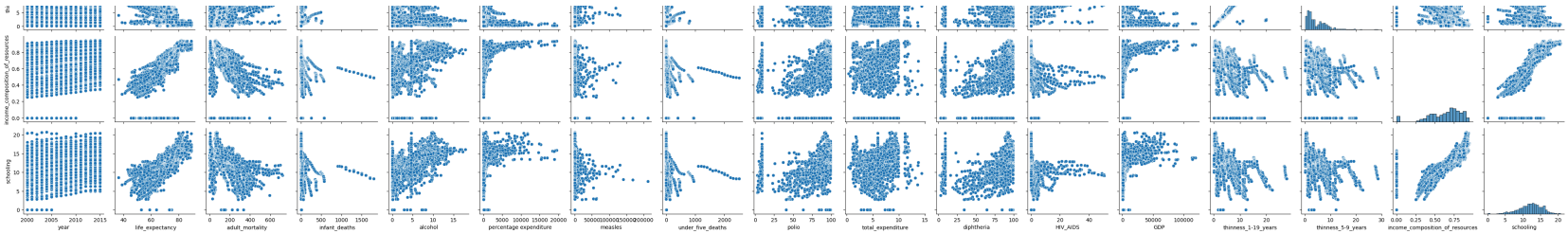
- Polio have the highes mean value
- HIV have the lowest mean value
- Polio also has the highes median value which gives it the highes values among all columns

In [255]:
```
num_cols = life_expectancy_df.select_dtypes(include=['float64', 'int64']).columns

sns.pairplot(life_expectancy_df[num_cols])
plt.suptitle('Life Expectancy Variables', y=1.02)
plt.show()
```

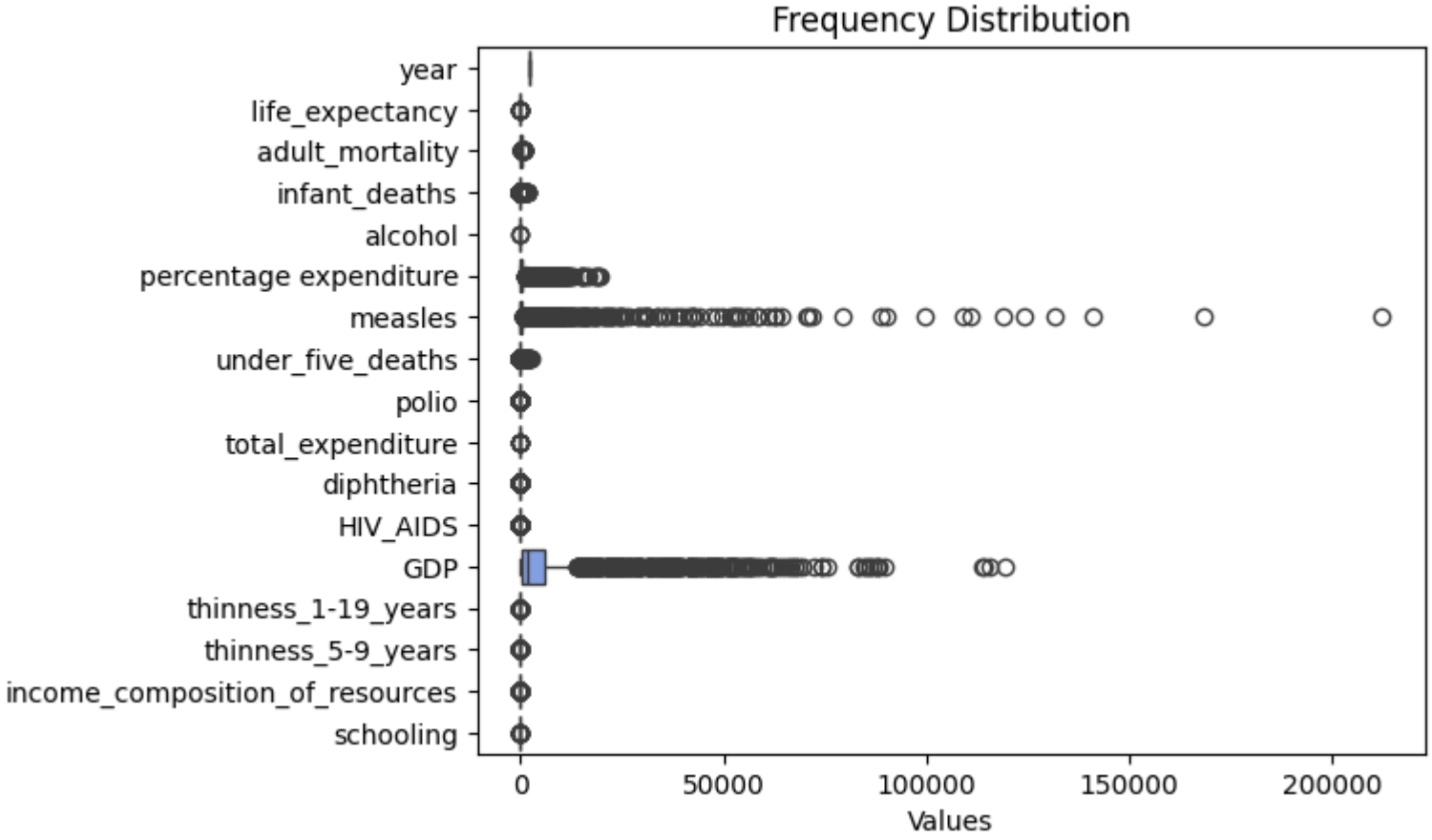
Life Expectancy Variables





In [259]:
```
sns.boxplot(data=life_expectancy_df, orient='h')

plt.title('Frequency Distribution')
plt.xlabel('Values')
plt.show()
```



In [ ]: