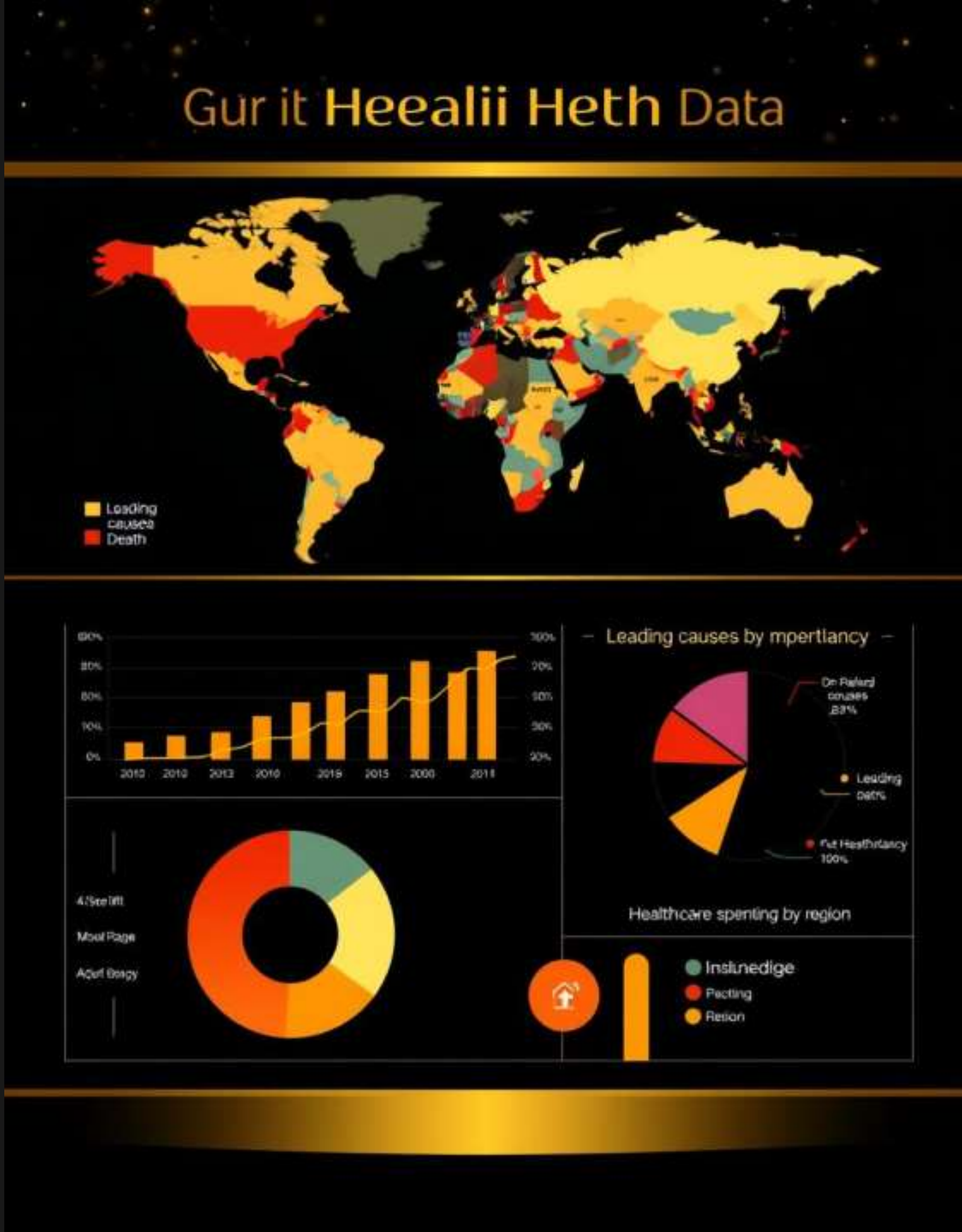
Life Expectancy Data Analysis: Project Overview

This project explores life expectancy trends using Global Health Observatory (GHO) data from a period of 2000 to 2015. Understanding these trends is crucial for addressing global health challenges

by NIMIT TIWARI



Problem Statement: Global Health Disparities

Significant variations in life expectancy exist across countries and regions. These disparities are often linked to socioeconomic factors.

This involves examining various factors that may affect life expectancy rates in different countries.

The project identifies key determinants influencing these differences.

- Countries Status : Developed or Developing countries
- Inconsistent Healthcare
- Poverty

Data Sources and Scope (2000-2023)

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries.

We have considered data from year 2000-2015 for 193 countries for further analysis.

The dataset consists of 22 Columns and 2938 rows which meant 20 predicting variables

All predicting variables are divided into several broad categories :

- Immunization related factors
- Mortality factors
- Economical factors
- Social factors

Methodology: Data Analysis Workflow

Our analysis follows a structured workflow. Each step contributes to actionable insights. This workflow ensures a thorough and data-driven approach

1. **Import Required Libraries and Load the Dataset**
   Importing libraries and exploring the dataset
2. **Data cleaning**
   Handling missing values, Outliers and inconsistencies
3. **EDA**
   Visualize data, identify pattern and correlation
4. **Statistical Analysis**
   Statistical analysis and hypothesis testing
5. **Model building and Evaluation**
   Linear regression, VIF and evaluation metric
6. **Key Findings and Conclusion**
   Result and prediction

Importing Required libraries

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use('ggplot')
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
from scipy import stats
from sklearn.linear_model import LogisticRegression
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn import metrics

import warnings
warnings.filterwarnings('ignore')
```

Exploring the dataset

```python
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
```

Data Cleaning

Exploring null values

```
df.isnull().sum()
```

```
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
```

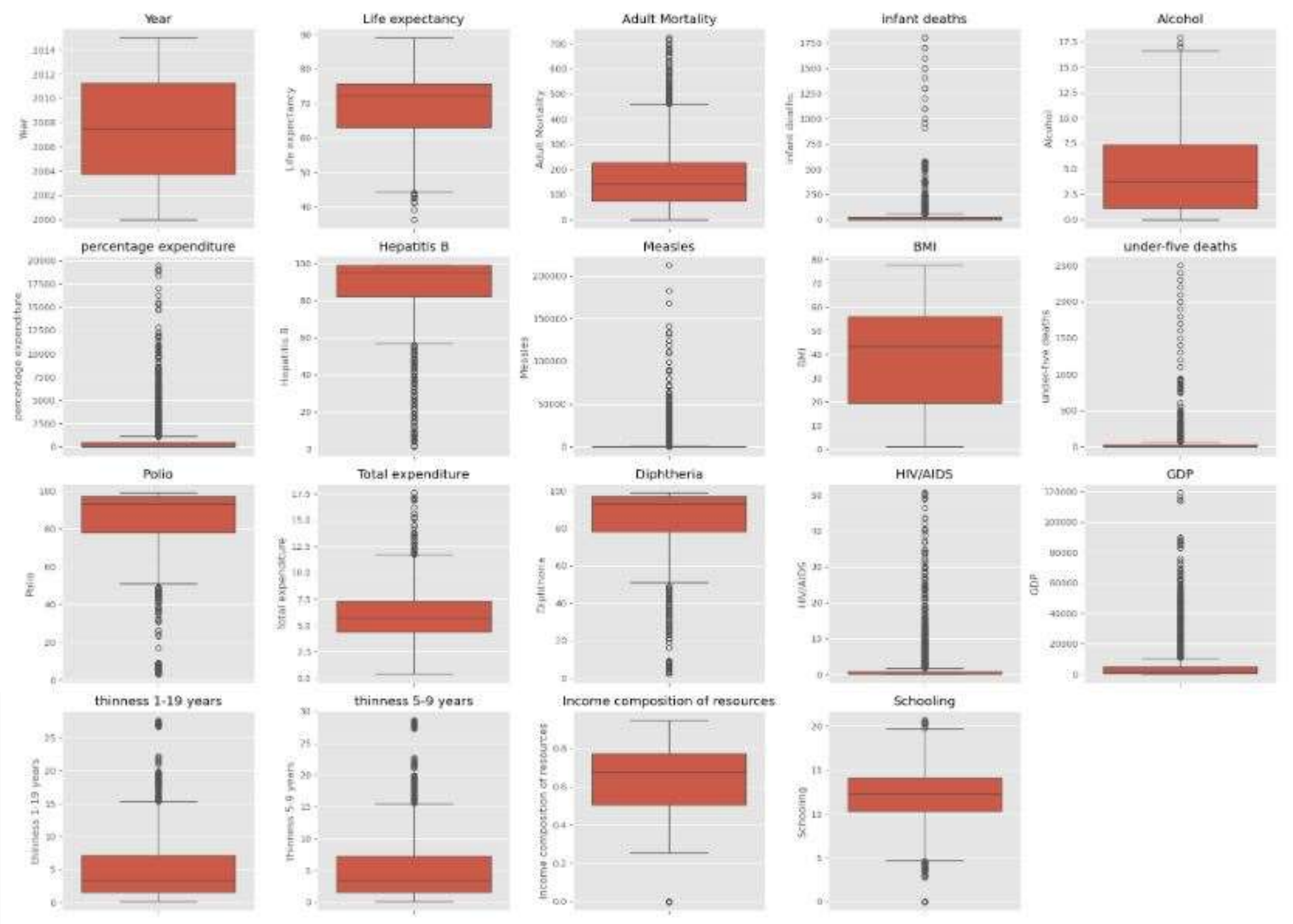
Imputing null values

```
# For numerical features with continous values: Impute with median (less sensitive to outliers)

num_cols_cont = ['Adult Mortality', 'Alcohol', 'BMI', 'Total expenditure','GDP',
                 'thinness 1-19 years', 'thinness 5-9 years','Income composition of resources', 'Schooling']

df1[num_cols_cont] = df1[num_cols_cont].fillna(df1[num_cols_cont].median())
```

```
# For numerical features with continous values: Impute with median (less sensitive to outliers)

num_cols_disc = ['Hepatitis B', 'Polio', 'Diphtheria']

df1[num_cols_disc] = df1[num_cols_disc].fillna(df1[num_cols_disc].mode().iloc[0])
```

Outlier

```python
fig, axes = plt.subplots(4, 5, figsize=(20, 16))
fig.suptitle('Boxplots of Numerical Columns', fontsize=16)

axes = axes.flatten()

for i, col in enumerate(numeric_columns):
    sns.boxplot(y=df1[col], ax=axes[i])
    axes[i].set_title(col)

# Remove any empty subplots
for j in range(len(numeric_columns), len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```



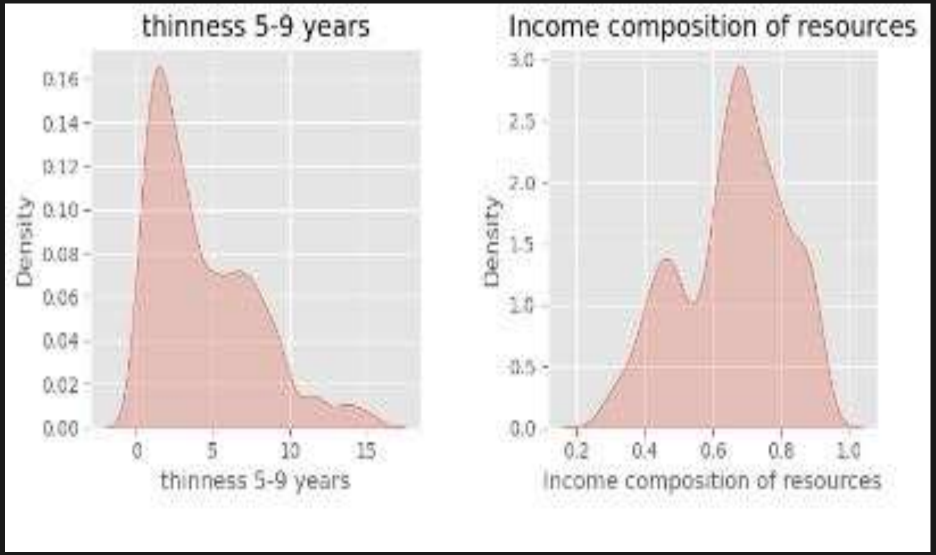
Removing outlier using IQR Method

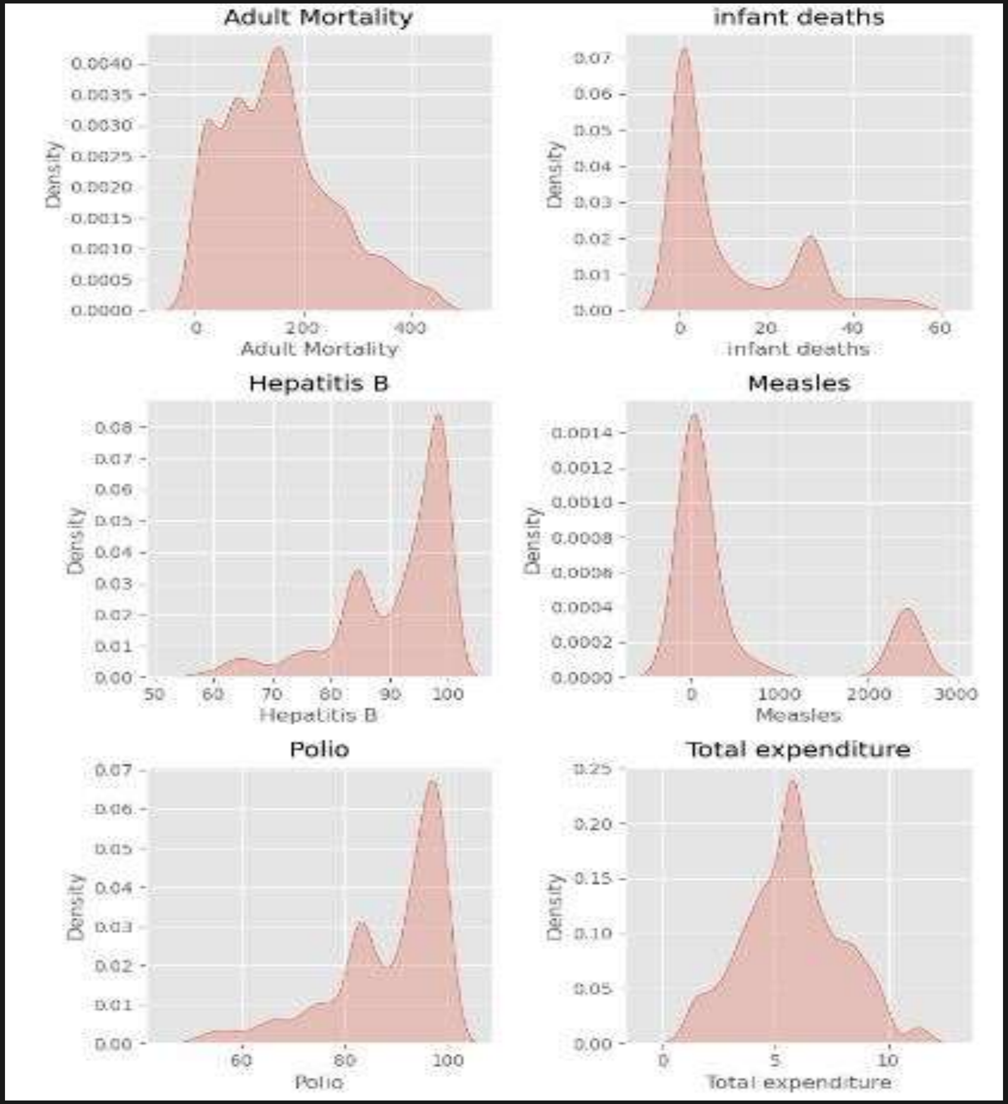
```python
for col_name in outlier_cols:

    q1 = df1[col_name].quantile(0.25)
    q3 = df1[col_name].quantile(0.75)
    iqr = q3 - q1

# The lower and upper bounds for outliers

    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr

# Replace outliers with the mean value of the column

    df1[col_name] = np.where((df1[col_name] > upper_bound) | (df1[col_name] < lower_bound),
                             np.mean(df1[col_name]),
                             df1[col_name])
```

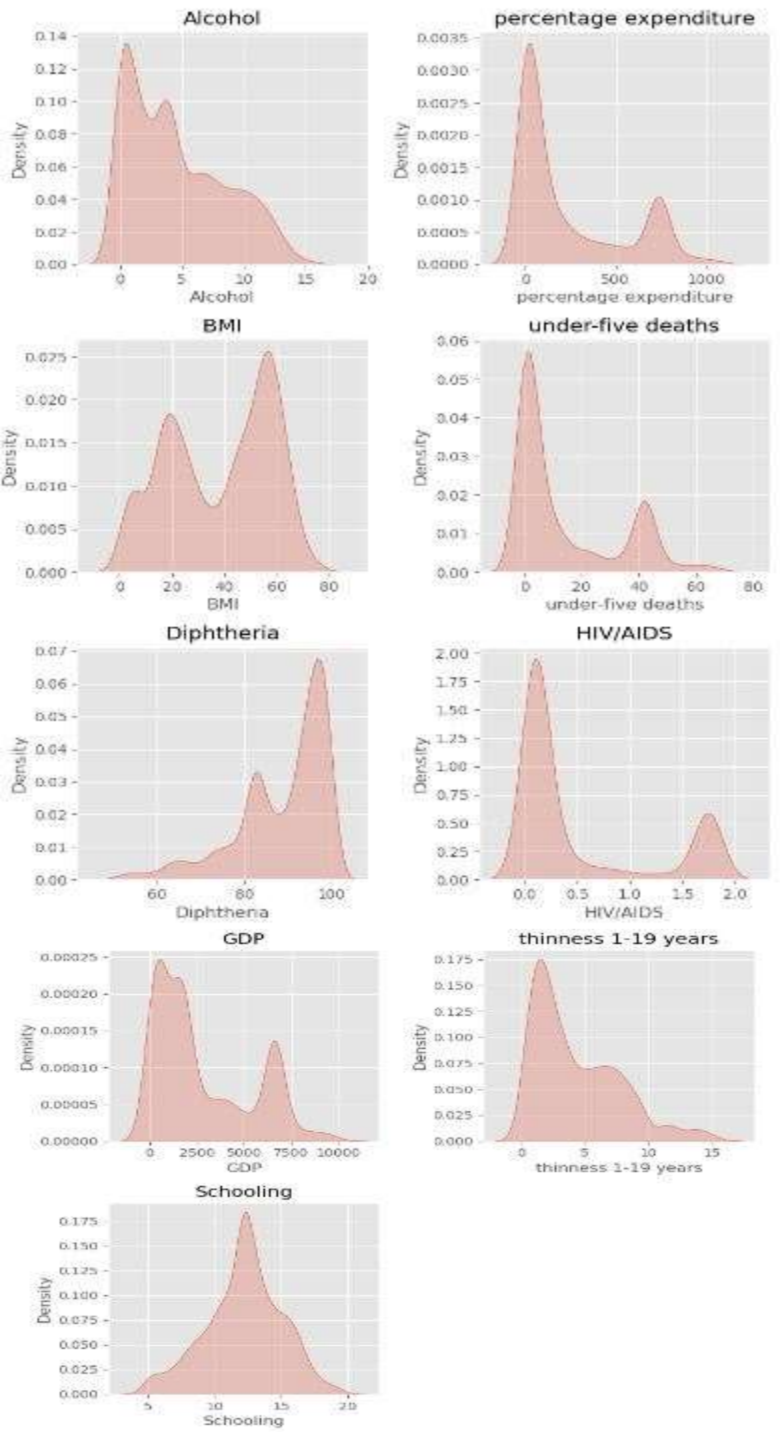
Exploratory Data Analysis (EDA)

EDA involved statistical summaries and visualizations

```python
# List of numerical columns (excluding Year and categoricals)
num_cols = df1.select_dtypes(include=['float64', 'int64']).columns.drop(['Year', 'Life expectancy'])

# Plot KDE for each column
plt.figure(figsize=(15, 20))
for i, col in enumerate(num_cols, 1):
    plt.subplot(6, 4, i)
    sns.kdeplot(df1[col], fill=True)
    plt.title(f'{col}')
plt.tight_layout()
plt.show()
```

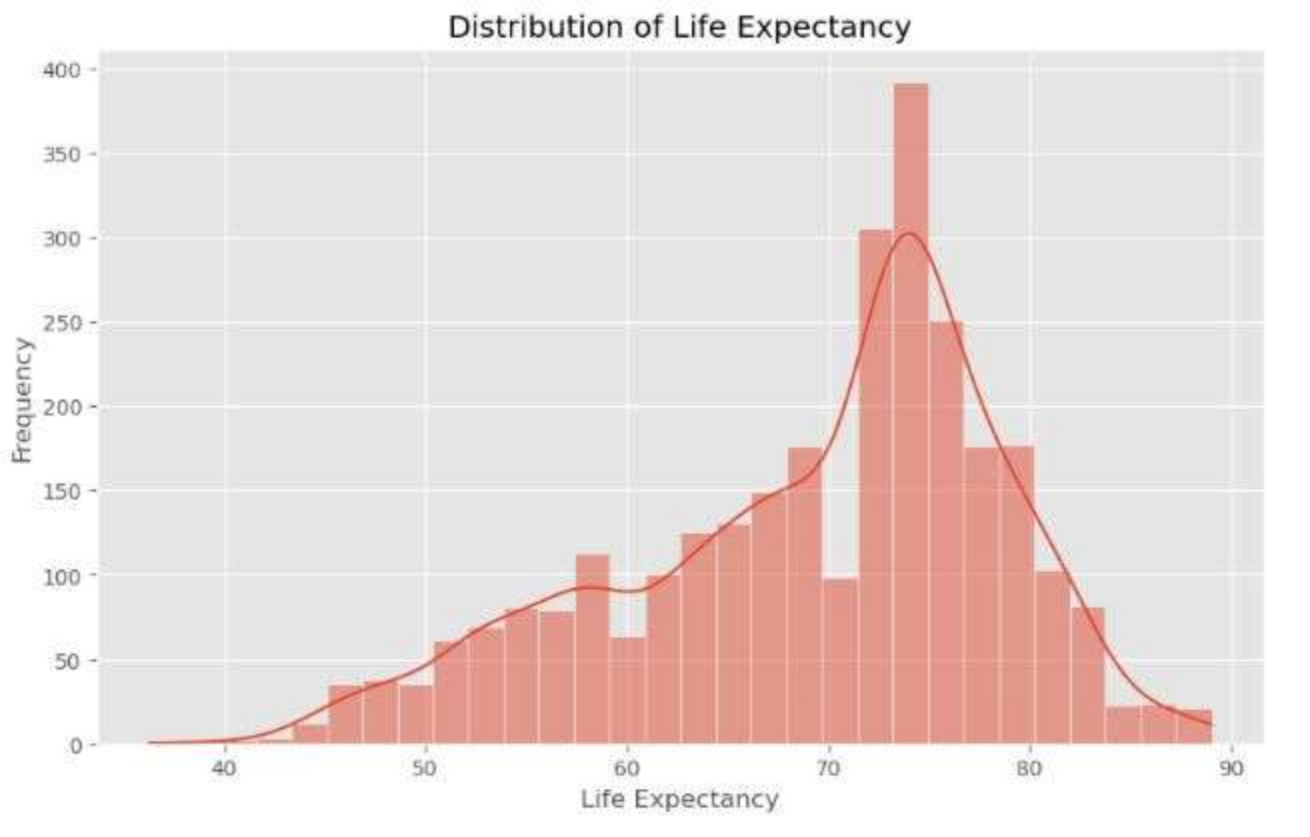






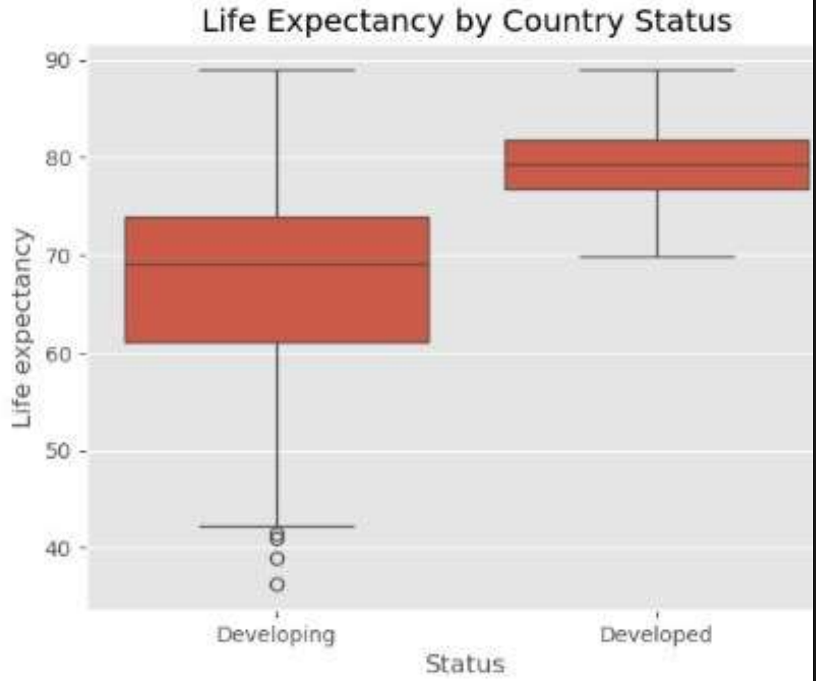
Target Variable Analysis (Life expectancy)

```
# Distribution of Life expectancy

plt.figure(figsize=(10, 6))
sns.histplot(df1['Life expectancy'], bins=30, kde=True)
plt.title('Distribution of Life Expectancy')
plt.xlabel('Life Expectancy')
plt.ylabel('Frequency')
plt.show()
```

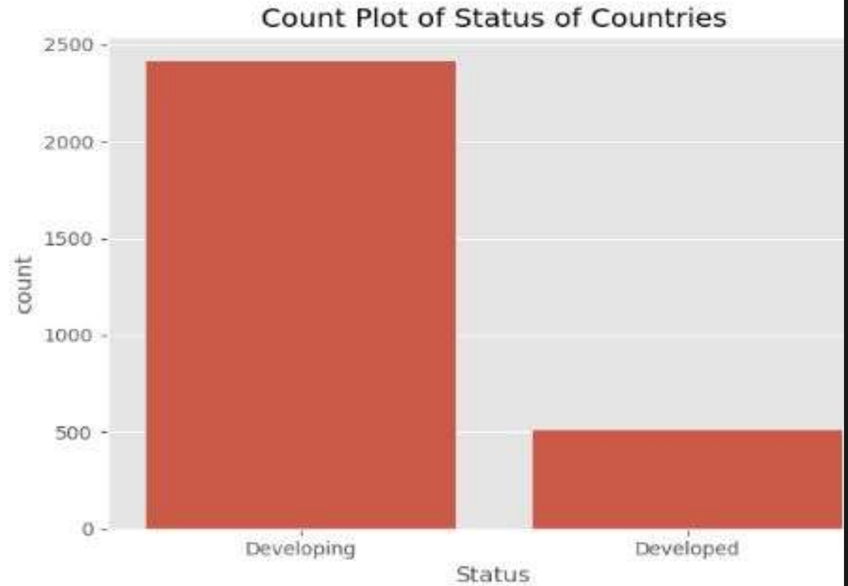


```python
# Boxplot by 'Status'

sns.boxplot(x='Status', y='Life expectancy', data=df1)
plt.title("Life Expectancy by Country Status")
```

```
Text(0.5, 1.0, 'Life Expectancy by Country Status')
```

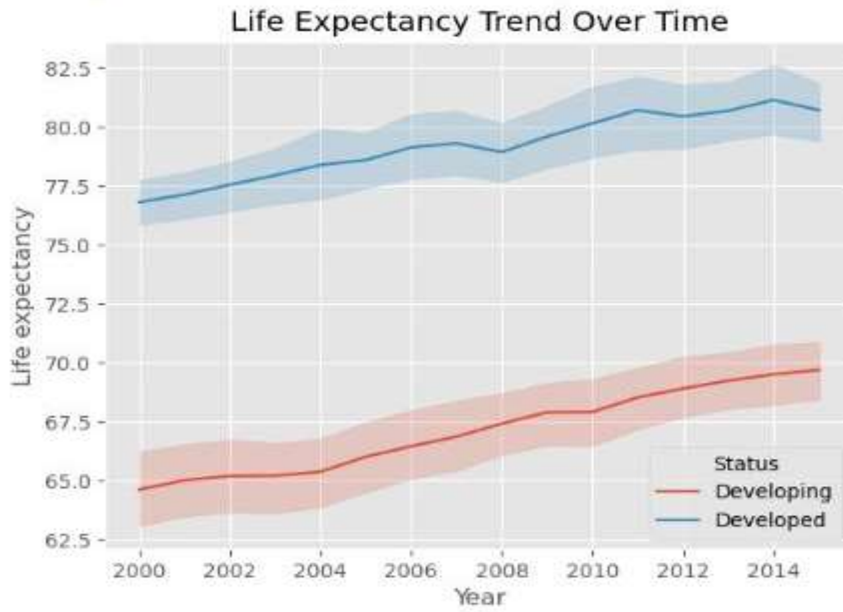



```python
sns.countplot(x=df1['Status'])
plt.title('Count Plot of Status of Countries')
plt.tight_layout()
plt.show()
```

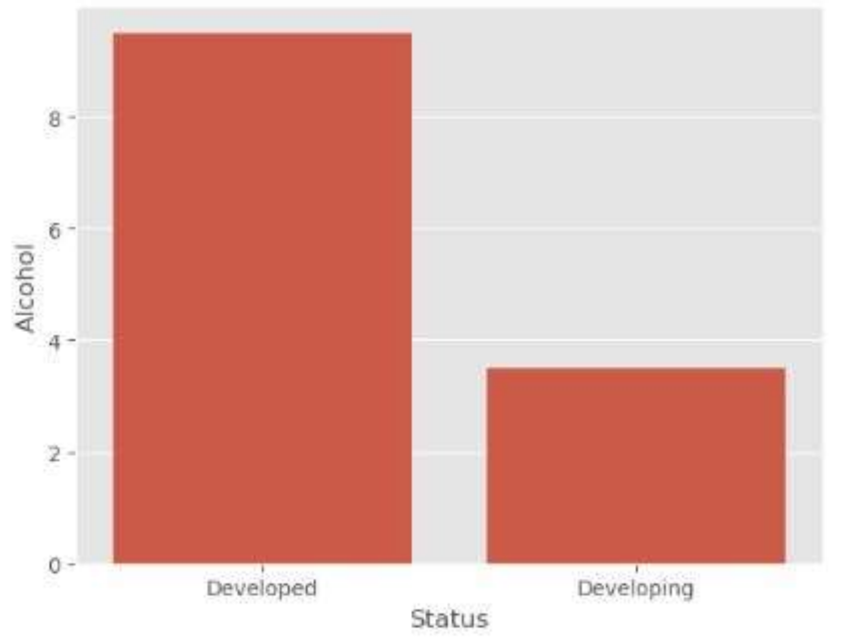



```python
sns.lineplot(x='Year', y='Life expectancy', hue='Status', data=df)
plt.title("Life Expectancy Trend Over Time")
```

```
Text(0.5, 1.0, 'Life Expectancy Trend Over Time')
```

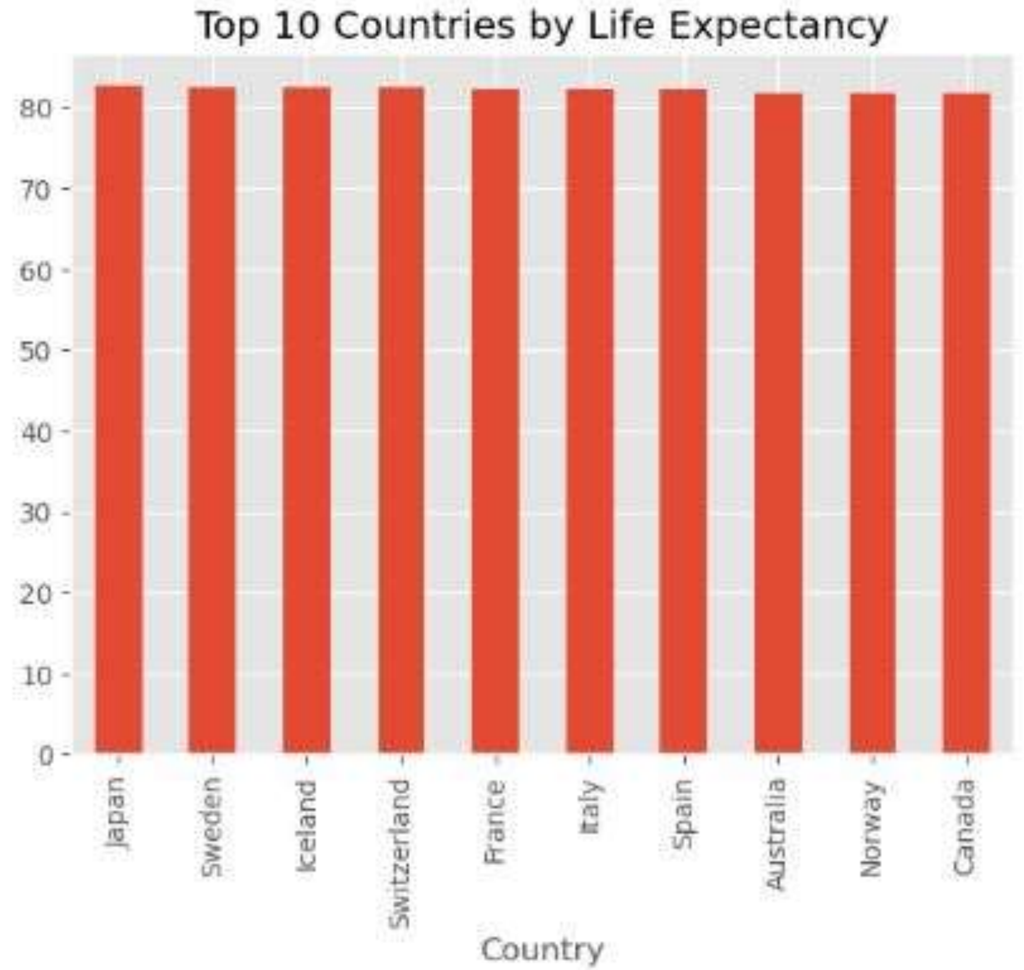



```python
data = df1.groupby('Status')['Alcohol'].mean().reset_index()

sns.barplot(data = data, x= 'Status', y= 'Alcohol')
plt.show()
```



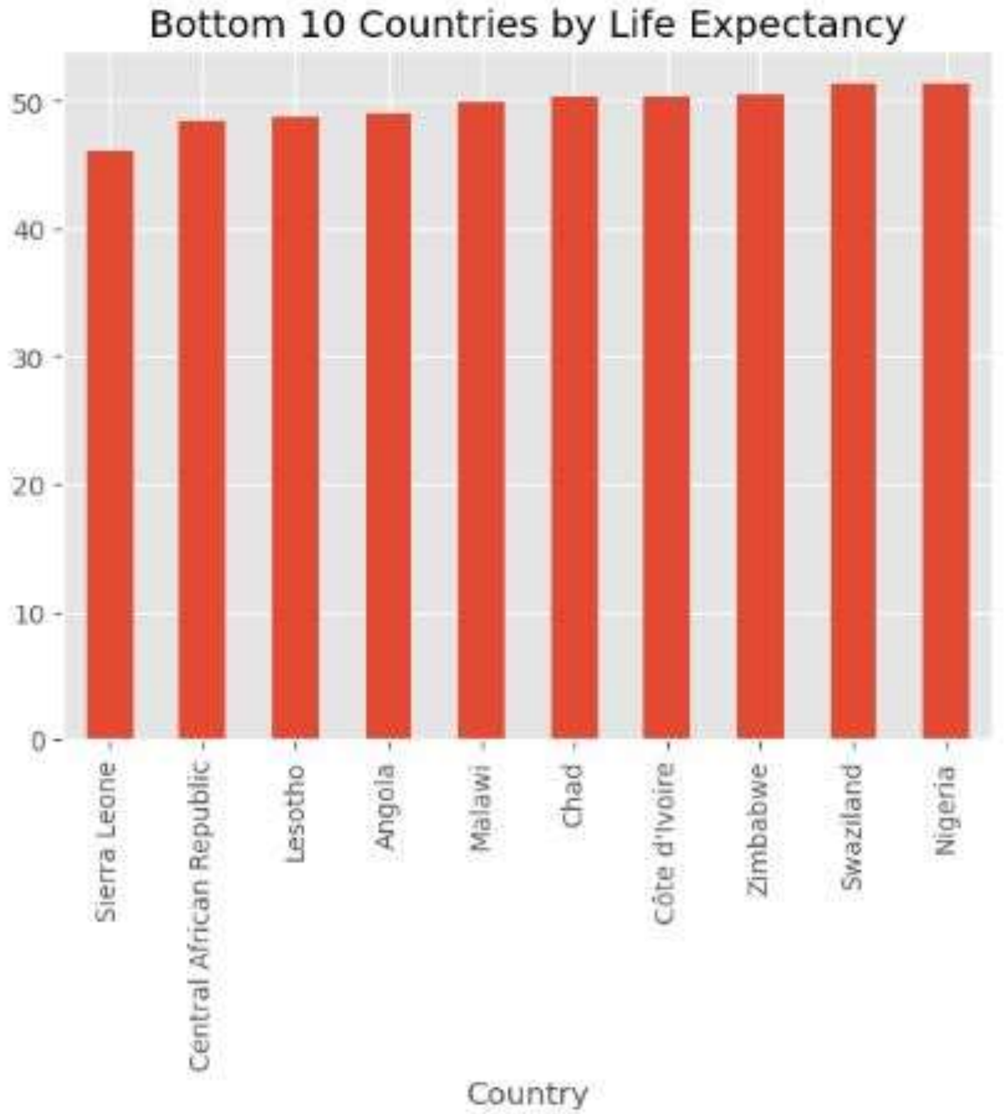
```python
top10 = df1.groupby('Country')['Life expectancy'].mean().nlargest(10)
bottom10 = df1.groupby('Country')['Life expectancy'].mean().nsmallest(10)

top10.plot(kind='bar', title="Top 10 Countries by Life Expectancy")

# bottom10.plot(kind='bar', title="Bottom 10 Countries by Life Expectancy")
```

```
<Axes: title={'center': 'Top 10 Countries by Life Expectancy'}, xlabel='Country'>
```



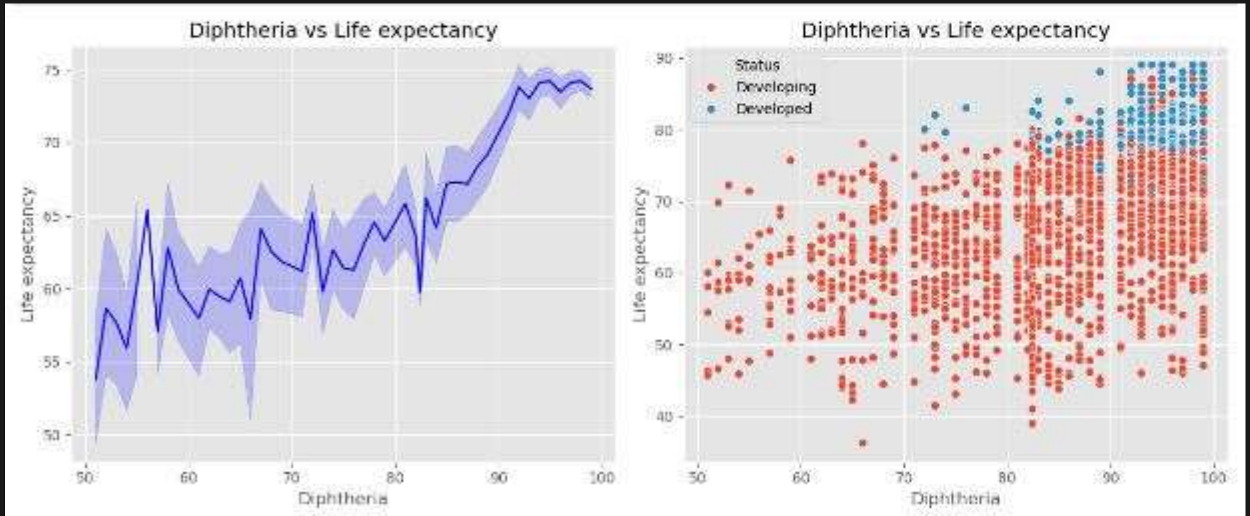

```python
bottom10.plot(kind='bar', title="Bottom 10 Countries by Life Expectancy")
```

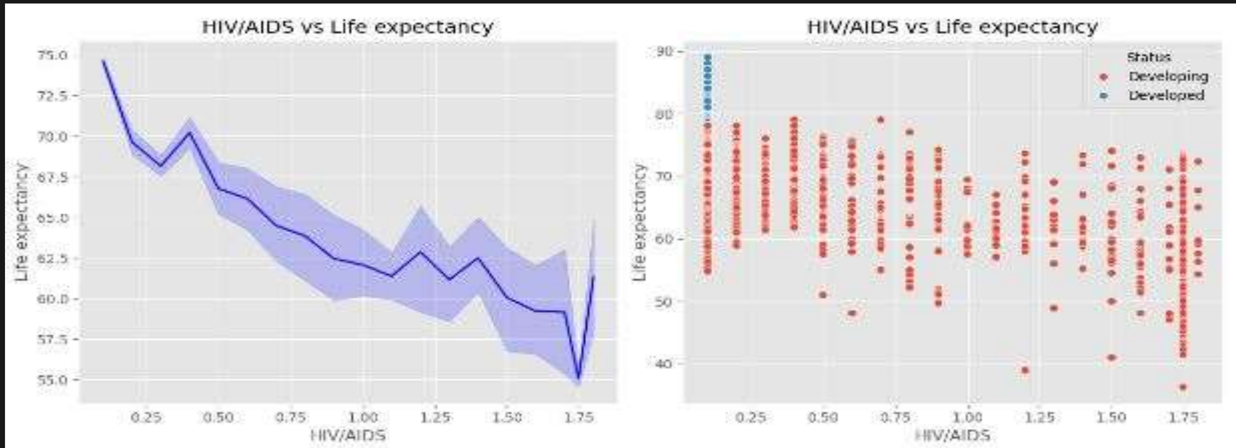
```
<Axes: title={'center': 'Bottom 10 Countries by Life Expectancy'}, xlabel='Country'>
```

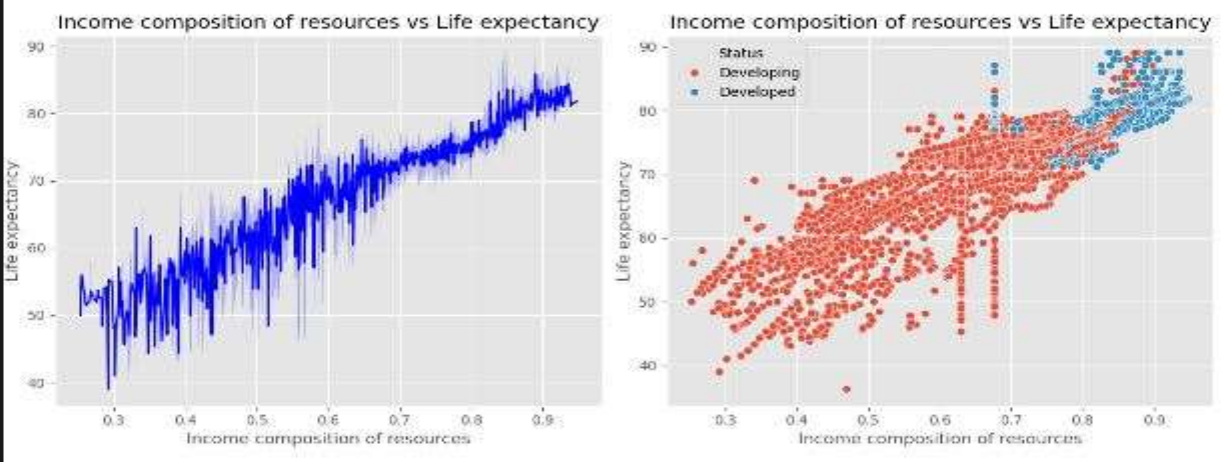


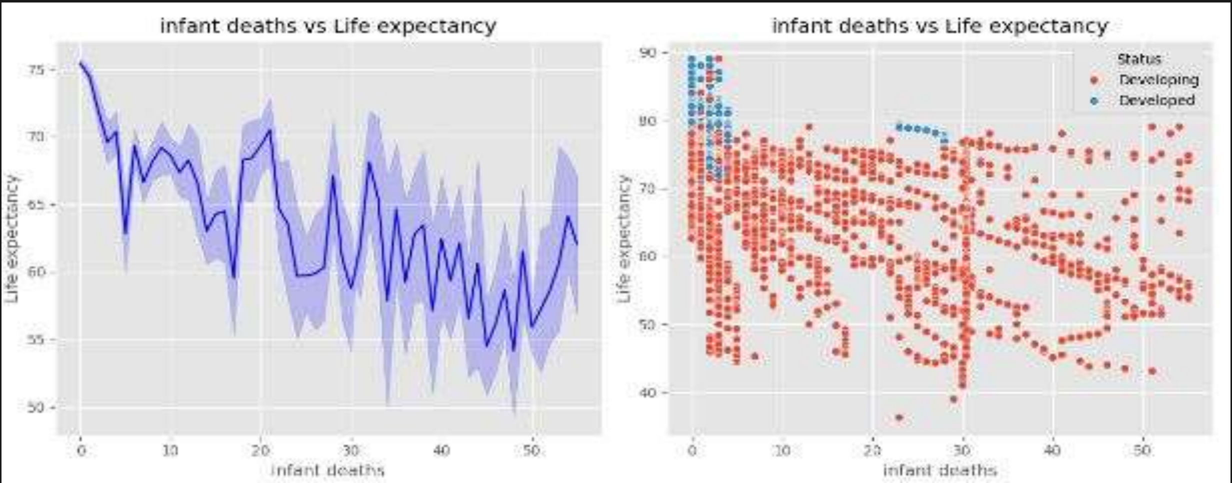
```python
def plot_line_and_scatter(x_vars):

    for x in x_vars:
        # Create a figure with two side-by-side subplots
        fig, (ax1, ax2) = plt.subplots(
            nrows=1,
            ncols=2,
            figsize=(12,5)
        )

        # Line plot
        sns.lineplot(data=df1, x=x, y='Life expectancy', ax=ax1, color='blue')
        ax1.set_title(f'{x} vs Life expectancy')

        # Scatter plot
        sns.scatterplot(data=df1, x=x, y='Life expectancy',hue='Status', ax=ax2, color='red')
        ax2.set_title(f'{x} vs Life expectancy')

        plt.tight_layout()
        plt.show()
```

```python
variables = ['Diphtheria', 'Polio','HIV/AIDS','Hepatitis B',
             'Schooling','Income composition of resources',
             'infant deaths','GDP','Adult Mortality','Total expenditure'
             ,'Alcohol','thinness 1-19 years','under-five deaths']
plot_line_and_scatter(x_vars=variables)
```

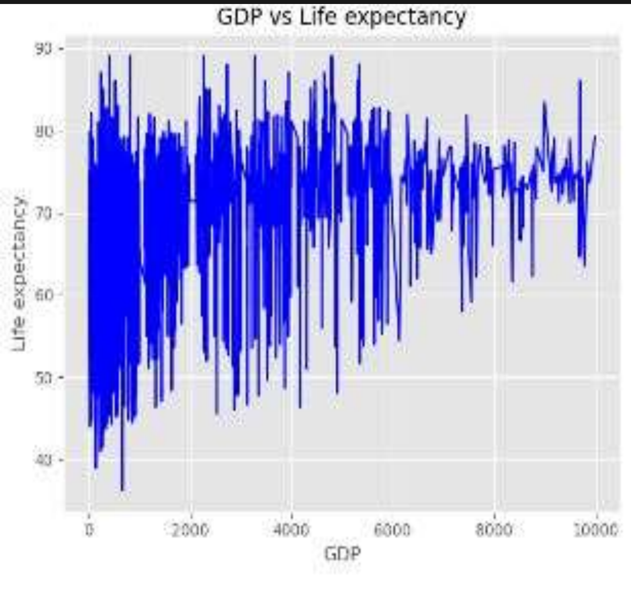


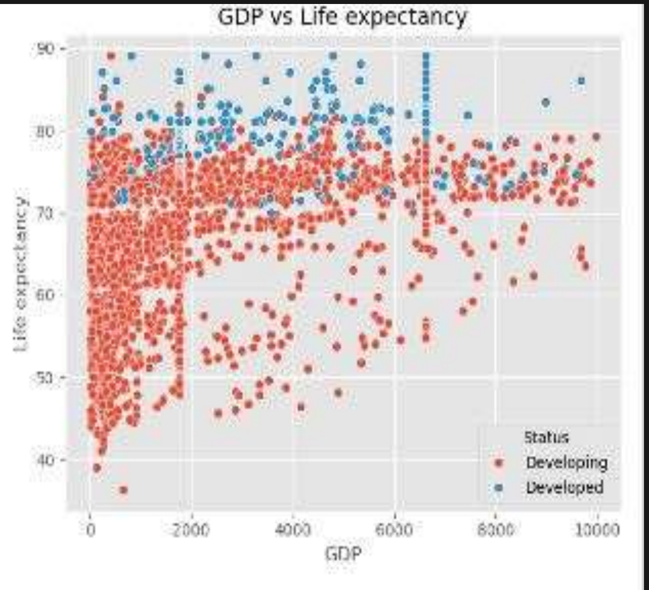


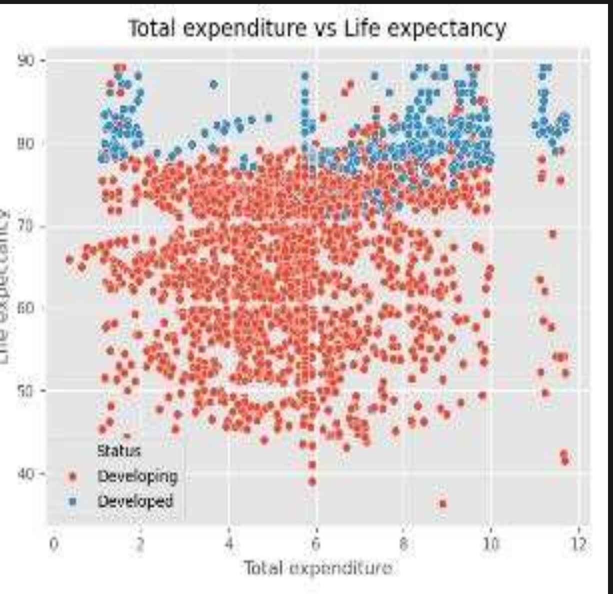


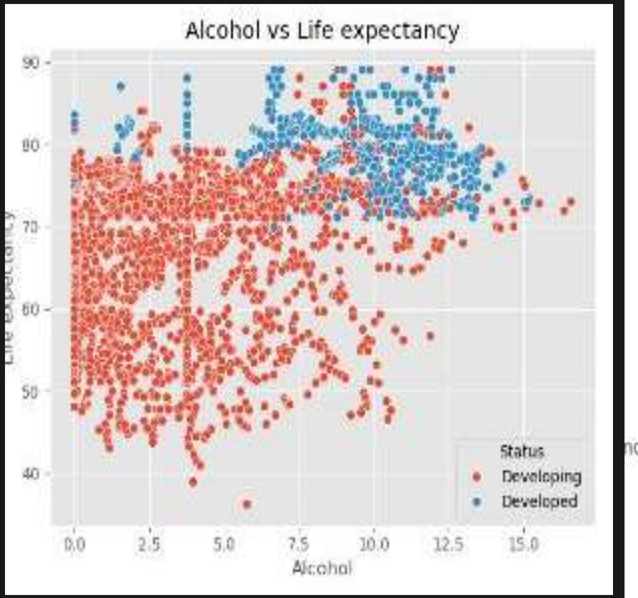


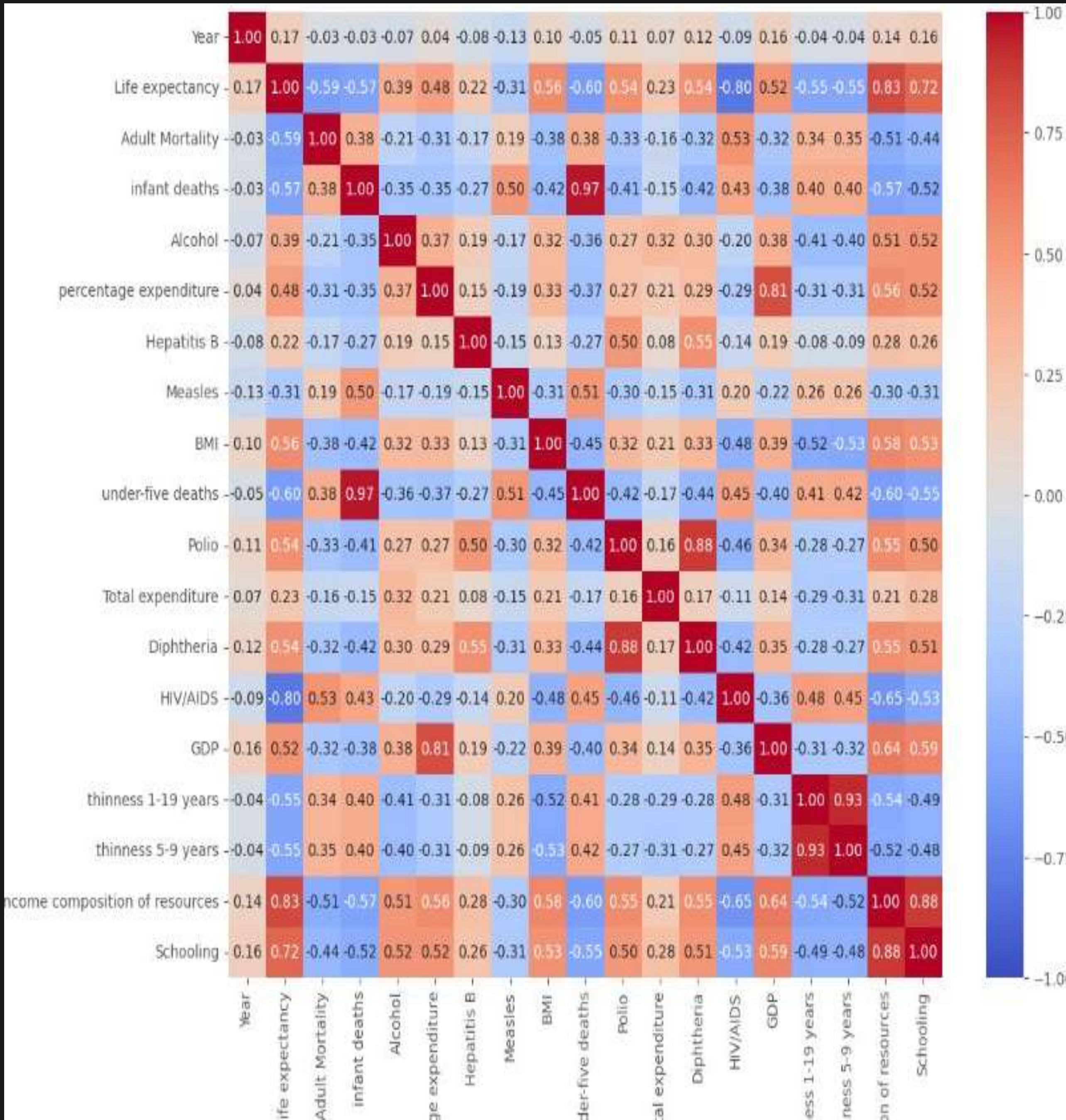
Correlation metric











Statistical Analysis

Correlation between life expectancy and key indicators

```python
# Calculate Pearson correlation matrix
corr_matrix = df1.corr(method='pearson')

# Focus on life expectancy correlations
life_expectancy_corr = corr_matrix['Life expectancy'].sort_values(ascending=False)

print("Top Correlations with Life Expectancy:")
print(life_expectancy_corr)
```

```
Top Correlations with Life Expectancy:
Life expectancy                    1.000000
Income composition of resources    0.827501
Schooling                          0.719856
BMI                                0.558888
Polio                              0.540304
Diphtheria                         0.538189
GDP                                0.524044
percentage expenditure             0.475629
Alcohol                            0.391483
Total expenditure                  0.229477
Hepatitis B                        0.215748
Year                               0.170033
Measles                           -0.313349
Status_Developing                 -0.482136
thinness 5-9 years                -0.547954
thinness 1-19 years               -0.552990
infant deaths                     -0.574799
Adult Mortality                   -0.594867
under-five deaths                 -0.600349
HIV/AIDS                          -0.796341
Name: Life expectancy, dtype: float64
```

Hypothesis Testing

Is there a significant difference in life expectancy between high-income and low-income countries?

```python
# # Using median GDP to split (adjust threshold as needed)
median_gdp = df1['GDP'].median()
df1['Income_Group'] = np.where(df1['GDP'] >= median_gdp, 'High', 'Low')

high_income = df1[df1['Income_Group'] == 'High']['Life expectancy']
low_income = df1[df1['Income_Group'] == 'Low']['Life expectancy']

t_stat, p_value = stats.ttest_ind(high_income, low_income)

print(f'T-test between High and Low Income Countries: t-statistic={t_stat:.2f},p-value={p_value:.2f}')

if p_value < 0.05:
    print("The difference in life expectancy between high-income and low-income countries is statistically significant.")
else:
    print("There is no statistically significant difference in life expectancy betweenhigh-income and low-income countries.")
```

t-statistic= 28.32 p-value= 0.00

Is there a significant difference in life expectancy between countries spending more or less on healthcare?

```python
median_expenditure = df1['percentage expenditure'].median()
df1['Expenditure_Group'] = np.where(df1['percentage expenditure'] >= median_expenditure, 'High', 'Low')

high_exp_life = df1[df1['Expenditure_Group'] == 'High']['Life expectancy']
low_exp_life = df1[df1['Expenditure_Group'] == 'Low']['Life expectancy']

t_stat, p_value = stats.ttest_ind(high_exp_life, low_exp_life)

print(f'T-test between High and Low health expenditure Countries: t-statistic={t_stat:.2f},p-value={p_value:.2f}')

if p_value < 0.05:
    print("The difference in life expectancy between high-health expenditure and low-health expenditure countries is statistically significant.")
else:
    print("There is no statistically significant difference in life expectancy between high-health expenditure and low-health expenditure countries.
```

t-statistic= 25.45 p-value= 0.00

Feature Scaling

```
numeric_columns = X_train.select_dtypes(include=['float64','int64']).columns

scaler = StandardScaler()

X_train[numeric_columns] = scaler.fit_transform(X_train[numeric_columns])
```

VIF

```
# Create a dataframe that will contain the names of all the feature variables and their respective VIFs
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in range(X_train.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

	Features	VIF
8	under-five deaths	19.75
2	infant deaths	18.27
16	Income composition of resources	7.59
14	thinness 1-19 years	7.49
15	thinness 5-9 years	7.41
11	Diphtheria	5.30
9	Polio	5.15
17	Schooling	4.92
13	GDP	3.62
4	percentage expenditure	3.17
12	HIV/AIDS	2.20
7	BMI	1.77
3	Alcohol	1.69
5	Hepatitis B	1.62
1	Adult Mortality	1.54
6	Measles	1.48
10	Total expenditure	1.24
0	Year	1.17
18	Status_Developing	1.09

Building a linear model

```
logm1 = sm.OLS(y_train,(sm.add_constant(X_train))).fit()
print(logm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     572.3
Date:                Wed, 16 Apr 2025   Prob (F-statistic):               0.00
Time:                        21:04:08   Log-Likelihood:                -5616.2
No. Observations:                2049   AIC:                         1.127e+04
Df Residuals:                    2029   BIC:                         1.138e+04
Df Model:                          19
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                              70.6911      0.257    275.364      0.000      70.188      71.195
Year                                0.6479      0.090      7.190      0.000       0.471       0.825
Adult Mortality                    -0.9624      0.104     -9.263      0.000      -1.166      -0.759
infant deaths                      -0.1615      0.356     -0.453      0.650      -0.860       0.537
Alcohol                            -0.0747      0.116     -0.643      0.520      -0.302       0.153
percentage expenditure              0.7354      0.148      4.954      0.000       0.444       1.027
Hepatitis B                        -0.2846      0.106     -2.682      0.007      -0.493      -0.077
Measles                            -0.1491      0.101     -1.472      0.141      -0.348       0.050
BMI                                -0.0176      0.111     -0.159      0.874      -0.235       0.200
under-five deaths                  -0.6364      0.371     -1.717      0.086      -1.363       0.091
Polio                              -0.2199      0.189     -1.163      0.245      -0.591       0.151
Total expenditure                   0.2646      0.093      2.839      0.005       0.082       0.447
Diphtheria                          0.8844      0.192      4.611      0.000       0.508       1.261
HIV/AIDS                           -3.5834      0.123    -29.034      0.000      -3.825      -3.341
GDP                                -0.4301      0.159     -2.712      0.007      -0.741      -0.119
thinness 1-19 years                 0.5038      0.228      2.210      0.027       0.057       0.951
thinness 5-9 years                 -0.9265      0.227     -4.083      0.000      -1.371      -0.481
Income composition of resources     3.9011      0.232     16.818      0.000       3.446       4.356
Schooling                          -0.4545      0.185     -2.460      0.014      -0.817      -0.092
Status_Developing                  -1.6617      0.295     -5.627      0.000      -2.241      -1.083
==============================================================================
Omnibus:                       73.652   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.224
Skew:                          -0.176   Prob(JB):                     2.43e-38
Kurtosis:                       4.380   Cond. No.                         16.6
==============================================================================
```

Rebuilding the model

```
X_train_sm = sm.add_constant(X_train)
logm_f = sm.OLS(y_train,X_train_sm).fit()
print(logm_f.summary())
```

	Features	VIF
8	Income composition of resources	3.67
6	GDP	3.56
2	percentage expenditure	3.09
5	HIV/AIDS	1.97
3	under-five deaths	1.73
1	Adult Mortality	1.52
4	Diphtheria	1.50
7	thinness 5-9 years	1.46
0	Year	1.07
9	Status_Developing	1.06

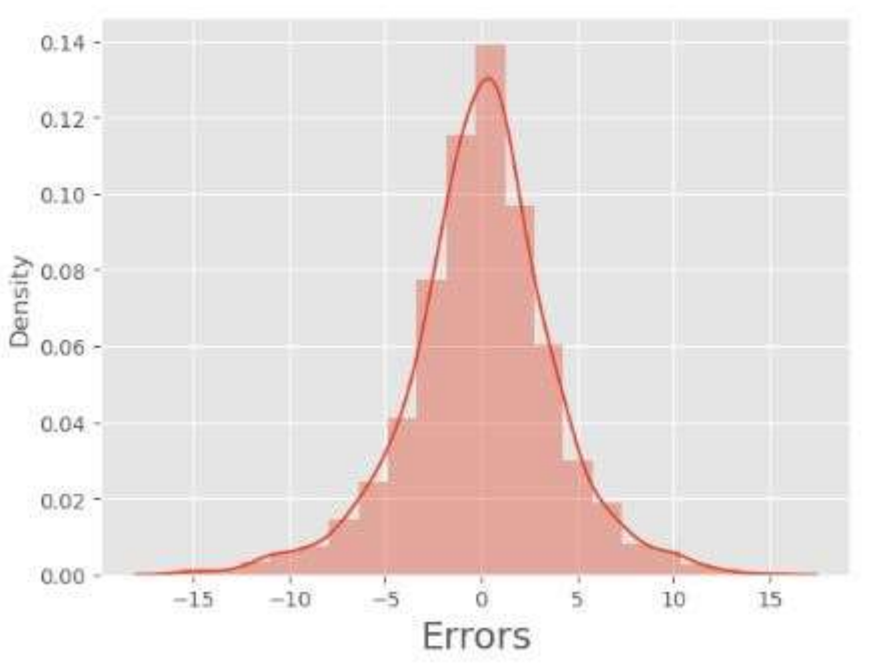
```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.841
Model:                            OLS   Adj. R-squared:                  0.840
Method:                 Least Squares   F-statistic:                     1075.
Date:                Wed, 16 Apr 2025   Prob (F-statistic):               0.00
Time:                        21:04:10   Log-Likelihood:                -5630.2
No. Observations:                2049   AIC:                         1.128e+04
Df Residuals:                    2038   BIC:                         1.134e+04
Df Model:                          10
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                              70.6294      0.236    299.870      0.000      70.168      71.091
Year                                0.7141      0.087      8.254      0.000       0.544       0.884
Adult Mortality                    -0.9879      0.103     -9.558      0.000      -1.191      -0.785
percentage expenditure              0.7916      0.147      5.371      0.000       0.503       1.081
under-five deaths                  -0.8318      0.110     -7.558      0.000      -1.048      -0.616
Diphtheria                          0.5549      0.102      5.418      0.000       0.354       0.756
HIV/AIDS                           -3.5853      0.118    -30.325      0.000      -3.817      -3.353
GDP                                -0.4818      0.158     -3.053      0.002      -0.791      -0.172
thinness 5-9 years                 -0.5602      0.103     -5.449      0.000      -0.762      -0.359
Income composition of resources     3.4174      0.167     20.434      0.000       3.089       3.745
Status_Developing                  -1.5868      0.268     -5.927      0.000      -2.112      -1.062
==============================================================================
Omnibus:                       75.292   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              172.902
Skew:                          -0.194   Prob(JB):                     2.85e-38
Kurtosis:                       4.369   Cond. No.                         8.71
==============================================================================
```

```
# Predicting the y_train

y_train_pred = logm_f.predict(X_train_sm)
```

```
res = y_train - y_train_pred
```

```
fig = plt.figure()
sns.distplot((res), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

Error Terms



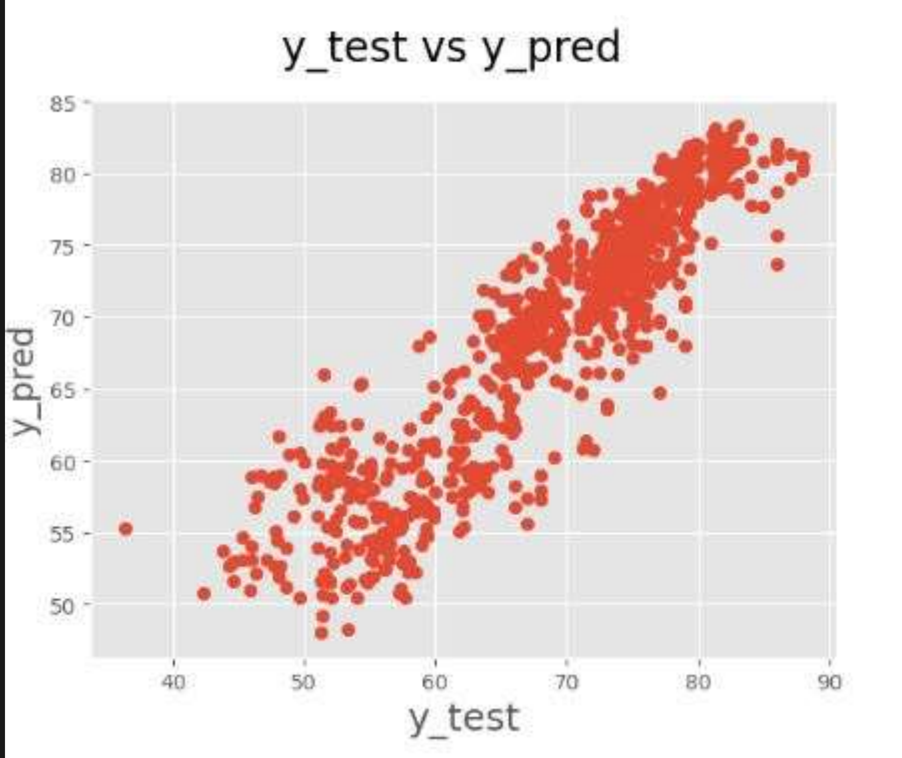
Making Predictions Using the Final Model

```
y_test_pred = logm_f.predict(X_test_sm)
```

```
mse = mean_squared_error(y_test, y_test_pred)
print(f"MSE: {mse:.2f}")

rmse = np.sqrt(mse)
print(f"RMSE: {rmse:.2f}")

r2 = r2_score(y_test, y_test_pred)
print(f"R²: {r2:.2f}")

mae = mean_absolute_error(y_test, y_test_pred)
print(f"MAE: {mae:.2f}")
```

```
MSE: 16.52
RMSE: 4.06
R²: 0.82
MAE: 3.03
```

```
# Plotting y_test and y_pred to understand the spread

fig = plt.figure()
plt.scatter(y_test, y_test_pred)
fig.suptitle('y_test vs y_pred', fontsize = 20)
plt.xlabel('y_test', fontsize = 18)
plt.ylabel('y_pred', fontsize = 16)
```



Key Findings and Observation

- The distribution of Life Expectancy have mean around 70 and median around 75.
- Developed countries exhibit significantly higher life expectancy compared to developing nations.
- The Life expectancy trends (2000–2014) for developed and developing countries increases over time for both groups at the same rate.
- The top 10 countries with the highest life expectancy (averaging 81.7–82.5 years) are all developed nations, led by: Japan (82.54 years),Sweden (82.52 years),Iceland (82.44 years) followed by Switzerland, France, Italy, Spain, Australia, Norway, and Canada.
- The bottom 10 countries for life expectancy (averaging 46.1–51.4 years) are all low-income African nations, with: Sierra Leone (46.1 years) at the lowest, Central African Republic (48.5 years), Lesotho (48.8 years), followed by Angola, Malawi, Chad, Côte d'Ivoire, Zimbabwe, Swaziland (Eswatini), and Nigeria.
- Analysis reveals a strong positive correlation between childhood immunization rates (diphtheria and polio) and national life expectancy figures. This relationship is particularly evident when comparing developed nations (with vaccination rates typically exceeding 90% and life expectancies above 80 years) against developing countries (where lower vaccination coverage correlates with life expectancies often below 70 years
- There is a strong positive correlation between a country's average years of schooling and its life expectancy. Countries with higher education levels (14+ years) typically have life expectancies above 75 years, while those with <12 years of schooling average <70 years.
- There is a strong positive correlation between a nation's income/resource composition and its life expectancy. Wealthier nations (higher GDP per capita) and those with equitable resource distribution consistently exhibit higher life expectancy.
- There is a strong negative correlation between HIV/AIDS prevalence and life expectancy. The data shows Low HIV/AIDS rates (0 deaths per 1,000 live births): Life expectancy ~75 years and High HIV/AIDS rates (2.0 deaths per 1,000 live births): Life expectancy drops sharply to ~55 years.
- A 1.0 increase in HIV/AIDS deaths (per 1,000 live births) corresponds with a ~10-year decline in life expectancy

Statistical Analysis and Model Result

Statistical Analysis Findings:

1.Income Level and Life Expectancy:

1. A **highly significant difference** exists in life expectancy between high-income and low-income countries ($t = 28.32$, $p < 0.001$).
2. **Interpretation:** Economic development strongly predicts longer lifespans, likely due to better healthcare infrastructure, nutrition, and sanitation in wealthier nations.

2.Healthcare Spending and Life Expectancy:

1. Countries with higher healthcare expenditures exhibit **significantly longer life expectancies** than those with lower spending ($t = 23.45$, $p < 0.001$).
2. **Interpretation:** Investment in healthcare systems directly correlates with improved population longevity, though efficiency and equitable access also play critical roles.

Top predictors include Income composition of resources, GDP, percentage expenditure, HIV/AIDS, under-five deaths, Adult Mortality, Diphtheria, thinness 5-9 years, Year, Status Developing.

Multiple Linear Regression : $R^2 = 0.84$.
RMSE = 4.02

Recommendations

- To improve overall life expectancy, we should focus more on developing countries through targeted healthcare investment and economic initiatives to push them toward becoming developed nations.
- If all nations matched Japan’s life expectancy, global averages would rise by ~5 years. Prioritizing preventive care and equitable healthcare access could close 50% of this gap.
- If the bottom 10 nations matched Rwanda’s progress (+10 years in 20 years), 50M+ lives could be saved per decade. This requires a large investment in targeted health aid.
- Closing the immunization gap in the 10 worst-performing nations could prevent more than 1M child deaths/year, adding ~5 years to their average life expectancy within a decades.
- A dual focus on education (especially for women) and equitable resource distribution could close the life expectancy gap between developed and developing nations.
- Eliminating mother-to-child HIV transmission in high-burden countries could increase national life expectancy

Thank You

N by Nimit Tiwari