

PRINCIPLES OF DATA SCIENCE  
ASDS5302 PROJECT FINAL REPORT

# **LIFE EXPECTANCY**

Project By:  
**SINDHURA SURAVARJHULA**  
**MRUNALI SAROJ NILGIRWAR**

Under the Guidance of:  
**Dr. MEI YANG**

## Contents

1. Introduction
2. Objective
3. Dataset Description
4. Methodology
5. Analysis
6. Conclusions
7. Results
8. Discussions

## **Introduction**

This project explores the factors affecting life expectancy around the world using regression techniques like Linear Regression, Ridge Regression, and Lasso Regression. The dataset includes information such as healthcare spending, mortality rates, and economic indicators for various countries over several years. By analyzing these factors, we aim to identify the key contributors to life expectancy and understand how they differ between developing and developed countries. The study provides insights that can help policymakers focus on improving public health outcomes through better resource allocation and planning.

## Objective

The main aim of this project is to analyze the factors that affect life expectancy in the world and develop predictive models to understand these relationships. Using demographic, health, and economic data, the project will compare the performance of multiple regression techniques to identify the most effective approach for predicting life expectancy.

### **Specific Goals:**

**Identification of Key Factors:** Use statistical and regression techniques to identify the most influential variables affecting life expectancy.

**Exploratory Data Analysis:** Investigate how life expectancy determinants differ across different variables, indicating where targeted improvement is necessary.

**Model Development and Comparison:** Develop a number of regression models including Linear Regression, Ridge Regression, and Lasso Regression, then compare the results to find the best fit.

**Validation and Insights:** Apply the models to unseen data to test their performance in terms of accuracy and generalization, and interpret the results to extract useful insights for policy decisions.

## **DATASET DESCRIPTION:**

The dataset used in this project is focused on factors influencing life expectancy across various countries and years. It includes both demographic and health-related variables, making it well-suited for understanding and predicting life expectancy. The dataset consists of 22 columns and contains data for multiple countries over several years.

### **Key Features**

1. **Demographics:**
  - **Country:** Name of the country.
  - **Year:** Year of the recorded data.
  - **Status:** Indicates whether the country is classified as 'Developed' or 'Developing'.
2. **Health Metrics:**
  - **Life Expectancy:** Target variable representing the average number of years a person is expected to live.
  - **Adult Mortality (per 1,000 adults):** Higher rates indicate poorer health outcomes.
  - **Infant Deaths:** Number of deaths of infants per 1,000 live births.
  - **Hepatitis B, Polio, Diphtheria:** Immunization coverage (% of the population).
3. **Lifestyle Factors:**
  - **Alcohol Consumption (per capita):** Levels of alcohol consumption by individuals.
  - **BMI:** Average Body Mass Index of the population.
  - **Thinness (1-19 years and 5-9 years):** Indicators of malnutrition.
4. **Economic and Resource Indicators:**
  - **GDP:** Gross Domestic Product per capita.
  - **Total Expenditure on Health (% of GDP):** Indicator of government spending on healthcare.
  - **Income Composition of Resources:** Index measuring human development factors like education and income.
  - **Schooling:** Average number of years of schooling.

### **Categorization of Features:**

1. **Numerical Predictors:**
  - Adult Mortality, Infant Deaths, Alcohol, BMI, Polio, Diphtheria, GDP, Thinness, Schooling, and others.
2. **Categorical Predictors:**
  - Status (Developed/Developing).
3. **Target Variable:**
  - Life Expectancy (continuous numerical variable).

## Methodology

The methodology for the analysis and prediction of life expectancy is systematic, ensuring that it is strong and accurate.

### 1. Data Preprocessing:

#### Missing Value Handling:

The missing values of continuous variables were imputed using the mean.

#### Outlier Detection and Treatment:

The outliers were discovered using boxplots and treated at the 1st and 99th percentiles to dampen their effect on the analysis.

#### Normalization:

Continuous variables have been scaled so that no model is influenced by the differences in scales.

#### Categorical Encoding:

Status variable, Developed/Developing, was one-hot encoded as the regression models cannot use categorical values directly.

### 2. Exploratory Data Analysis (EDA):

#### Distribution Analysis:

Firstly, histograms and boxplots were drawn for variables such as life expectancy, GDP, and health expenditure, among others.

#### Relationship Analysis:

Scatterplots were drawn to check for relationships between life expectancy and its predictors, namely adult mortality and GDP.

A correlation heatmap highlighted both strong and weak relationships between numerical variables.

#### Key Observations:

Life expectancy was negatively correlated with adult mortality and infant deaths.

On the other hand, life expectancy positively correlated with GDP, healthcare expenditure, and years of schooling.

### **3. Feature Selection**

Recursive Feature Elimination:

Features were ranked in order of their contribution to the predictive power for life expectancy.

LASSO Regression

Regularization helped identify important predictors and simultaneously removed less important variables to reduce noise.

### **4. Model Development**

Baseline Model:

Linear Regression was the baseline method for life expectancy prediction.

Advanced Models:

Ridge Regression and LASSO Regression were used to improve the prediction performance of the model and to handle the multicollinearity issue.

Comparison of Models:

The models are compared, and the best approach has been identified on the bases of predictive accuracy and interpretability.

### **5. Model Evaluation:**

Performance Metrics:

Performance metrics that have been used to assess the model include MAE, RMSE, and

R<sup>2</sup>-score.

Validation:

All models were validated on a hold-out test set to ensure generalizability.

Insights:

Comparison of each model was done, and the best model was interpreted for key factors influencing life expectancy.

# Analysis

## Step 1: Collect Data

### Data Description:

	Country	Year	Status	Life expectancy	Adult Mortality	Infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	1	65.0	263.0	62	0.01	71.279620	65.0	1154	...	6.0	8.16	65.0	0.1	58425920	33736494.0	17.2	17.3	0.479	10.1
1	Afghanistan	2014	1	59.9	271.0	64	0.01	73.523580	62.0	492	...	58.0	8.18	62.0	0.1	61269650	327582.0	17.5	17.5	0.476	10.0
2	Afghanistan	2013	1	59.9	268.0	66	0.01	73.219240	64.0	430	...	62.0	8.13	64.0	0.1	63174500	31731688.0	17.7	17.7	0.470	9.9
3	Afghanistan	2012	1	59.5	272.0	69	0.01	78.184220	67.0	2787	...	67.0	8.52	67.0	0.1	66995900	3696958.0	17.9	18.0	0.463	9.8
4	Afghanistan	2011	1	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.53723	2978599.0	18.2	18.2	0.454	9.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	1	44.3	723.0	27	4.36	0.000000	68.0	31	...	67.0	7.13	65.0	33.6	45436670	12777511.0	9.4	9.4	0.407	9.2
2934	Zimbabwe	2003	1	44.5	715.0	26	4.06	0.000000	7.0	998	...	7.0	6.52	68.0	36.7	45335120	12633897.0	9.8	9.9	0.418	9.5
2935	Zimbabwe	2002	1	44.8	73.0	25	4.43	0.000000	73.0	304	...	73.0	6.53	71.0	39.8	5734834	125525.0	1.2	1.3	0.427	10.0
2936	Zimbabwe	2001	1	45.3	686.0	25	1.72	0.000000	76.0	529	...	76.0	6.16	75.0	42.1	54858730	12366165.0	1.6	1.7	0.427	9.8
2937	Zimbabwe	2000	1	46.0	665.0	24	1.68	0.000000	79.0	1483	...	78.0	7.10	78.0	43.5	54735890	12222251.0	11.0	11.2	0.434	9.8

2938 rows x 22 columns

Data was imported in the ipynb file using the functions for analysis purpose.

### About the Data

The dataset contains detailed information about various factors influencing life expectancy, categorized into demographics, health indicators, and lifestyle attributes. The data provides insights into the socio-economic and healthcare variables that impact life expectancy globally. Below is a detailed description of the dataset's features:

#### Demographics:

- **Country:** Name of the country.
- **Year:** The year in which the data was recorded.
- **Status:** Indicates whether the country is classified as 'Developed' or 'Developing.'

#### Health Indicators:

- **Life Expectancy:** The average number of years a person is expected to live (Target Variable).
- **Adult Mortality:** Adult mortality rate per 1,000 population.
- **Infant Deaths:** The number of infant deaths per 1,000 live births.
- **BMI:** Average Body Mass Index of the population.
- **Hepatitis B, Polio, Diphtheria:** Immunization rates (%) against these diseases.
- **HIV/AIDS:** Death rate per 1,000 population due to HIV/AIDS.
- **Thinness (1-19 years):** Percentage of the population aged 1-19 who are underweight.
- **Thinness (5-9 years):** Percentage of the population aged 5-9 who are underweight.

#### Economic Indicators:

- **GDP:** Gross Domestic Product per capita.
- **Total Expenditure:** Total expenditure on health as a percentage of GDP.



- **Income Composition of Resources:** A measure of human development, including education and income.
- **Schooling:** Average number of years of schooling.

Lifestyle Indicators:

- **Alcohol:** Alcohol consumption per capita (liters of pure alcohol).


14

	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI ...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources
count	2938.000000	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	... 2919.000000	2712.000000	2919.000000	2938.000000	2490.000000	2.286000e+03	2904.000000	2904.000000	2771.000000
mean	2007.510720	0.825732	69.224932	164.796448	30.303948	4.602861	738.251288	80.940461	2419.592240	38.321247	... 82.550188	5.93819	82.324084	1.742103	7483.158519	1.275247e+07	4.839704	4.870317	0.627
std	4.613841	0.379405	9.523867	124.292079	117.926501	4.052413	1987.914824	25.070016	11467.272489	20.044034	... 23.428046	2.49832	23.716912	5.077785	14270.169394	6.096463e+07	4.420195	4.508882	0.211
min	2000.000000	0.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	... 3.000000	0.37000	2.000000	0.100000	1.681350	3.400000e+01	0.100000	0.100000	0.000
25%	2004.000000	1.000000	63.100000	74.000000	0.000000	0.877500	4.685342	77.000000	0.000000	19.300000	... 78.000000	4.26000	78.000000	0.100000	463.935625	1.957932e+05	1.600000	1.500000	0.497
50%	2008.000000	1.000000	72.100000	144.000000	3.000000	3.755000	64.912905	92.000000	17.000000	43.500000	... 93.000000	5.75500	93.000000	0.100000	1766.947500	1.386542e+06	3.300000	3.300000	0.677
75%	2012.000000	1.000000	75.700000	228.000000	22.000000	7.702500	441.534125	97.000000	360.250000	56.200000	... 97.000000	7.49250	97.000000	0.800000	5910.806250	7.420359e+06	7.200000	7.200000	0.777
max	2015.000000	1.000000	89.000000	723.000000	1800.000000	17.870000	19479.910000	99.000000	212183.000000	87.300000	... 99.000000	17.60000	99.000000	50.600000	119172.700000	1.290000e+09	27.700000	28.600000	0.941

8 rows × 21 columns

## Step 2: Data Cleaning

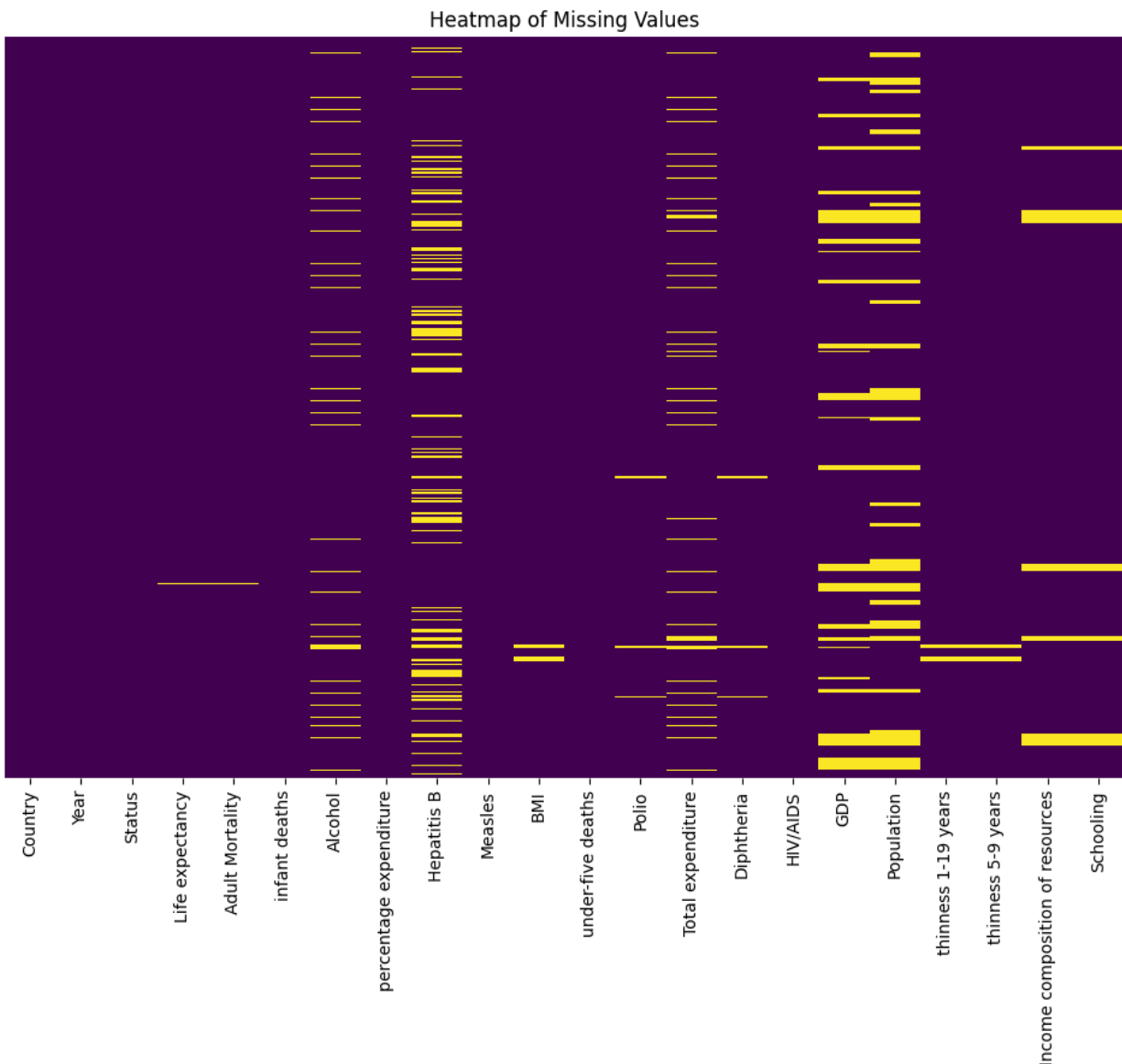
Number of Null Values for each column:



	0
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

Heatmap of missing values:



For each feature, the null values were checked, columns like Life Expectancy, Adult Mortality, Alcohol, Hepatitis B, BMI, Polio, Total Expenditure, Diphtheria, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling were found to be having null values, further **mean imputation** was done to proceed further with the analysis.

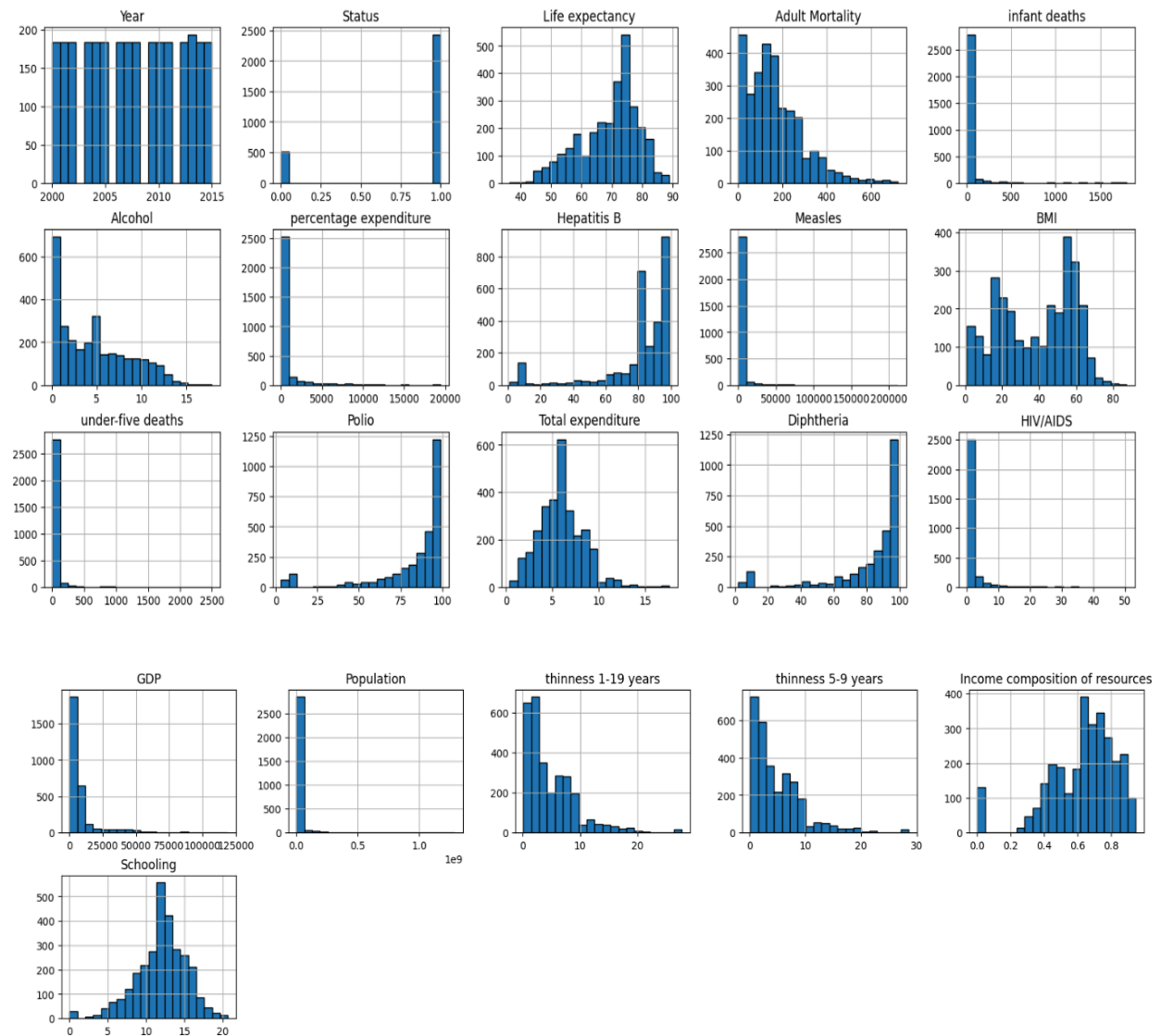
### Why Mean Imputation?

It is a common method for handling missing data in a large dataset. Especially, like in our dataset, where the data is missing completely at random, the estimate of mean remains unbiased.

## Step 3: Exploratory Data Analysis / Feature Engineering

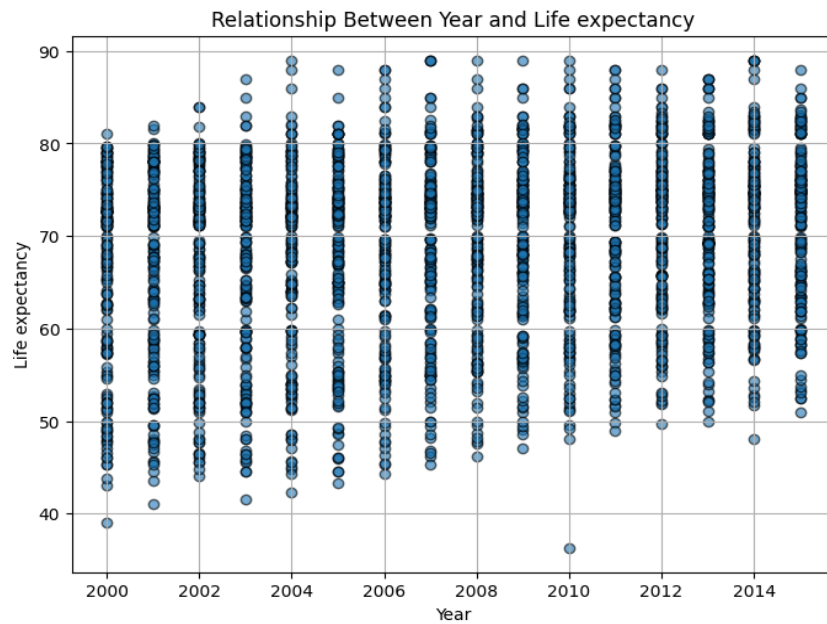
### Visualization:

Histograms of numeric variables:

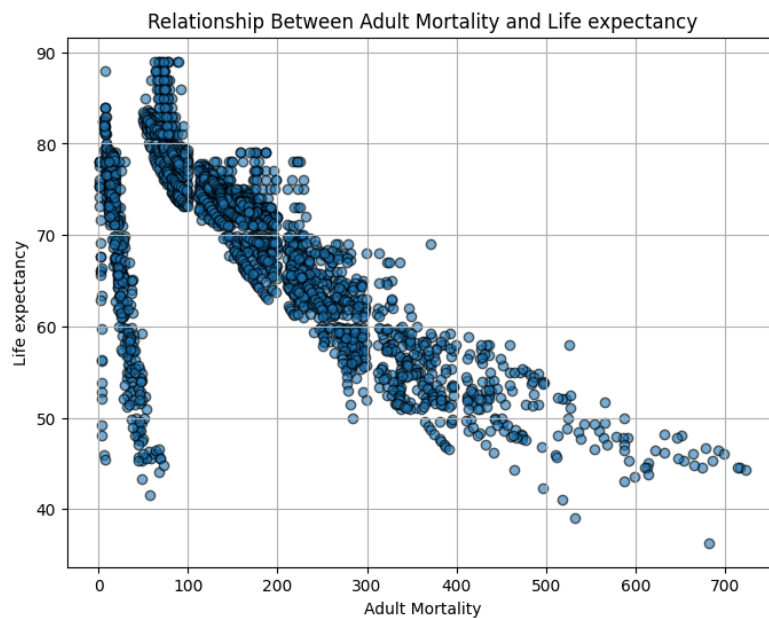


It can be observed that histograms are having varying distributions. None of them are uniformly distributed.

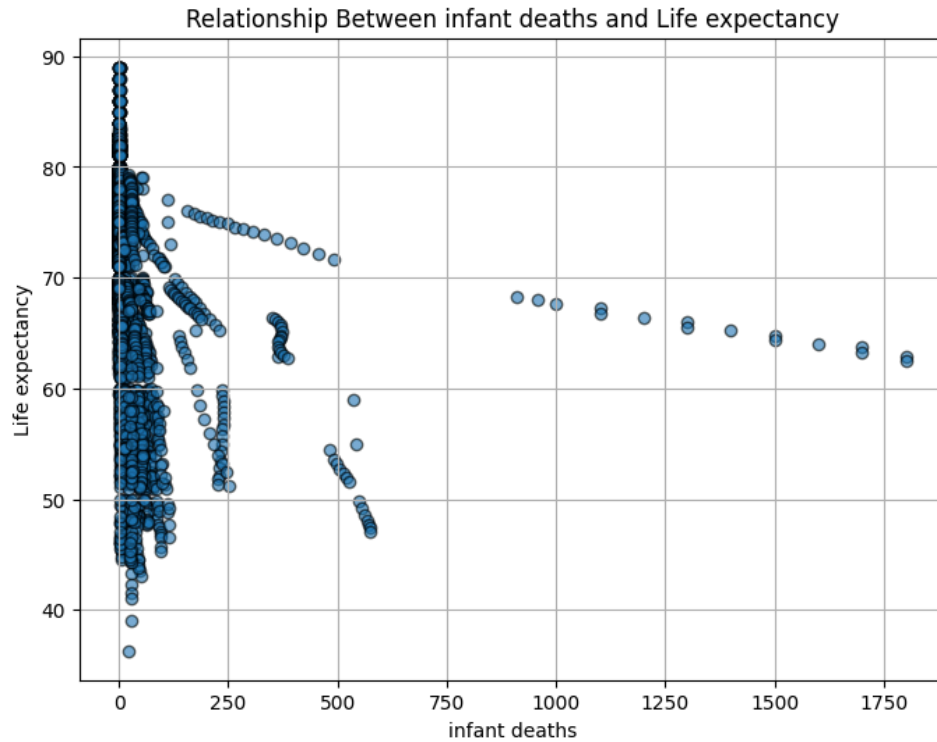
Created scatter plots of numerical variables against the target variable to observe the relationships between them.



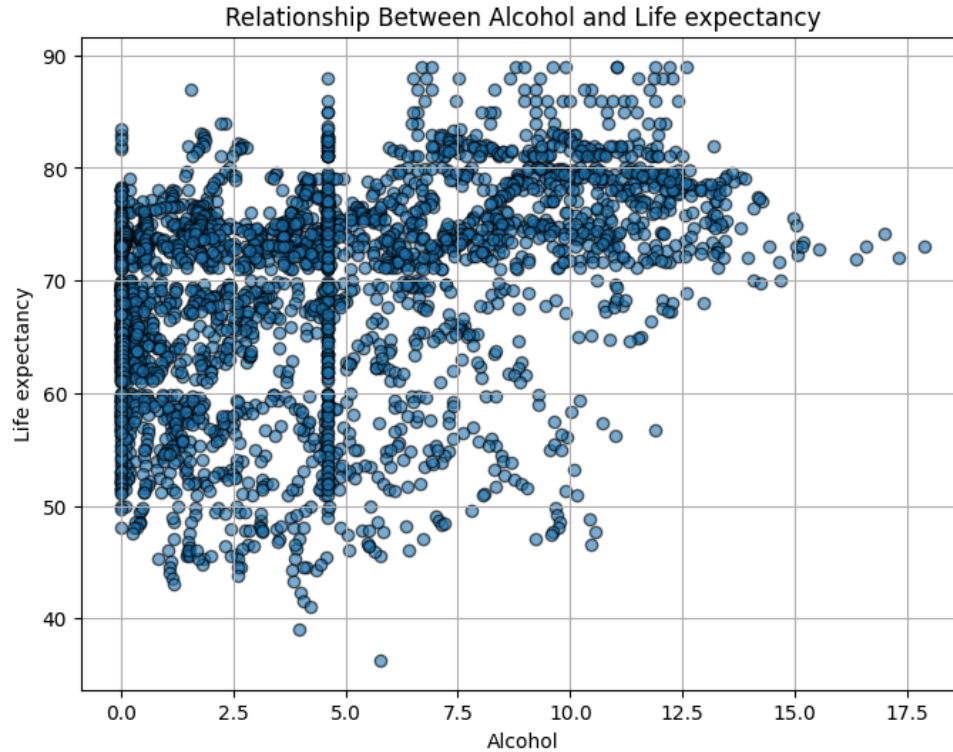
Life expectancy shows a general upward trend over the years, reflecting global health improvements.



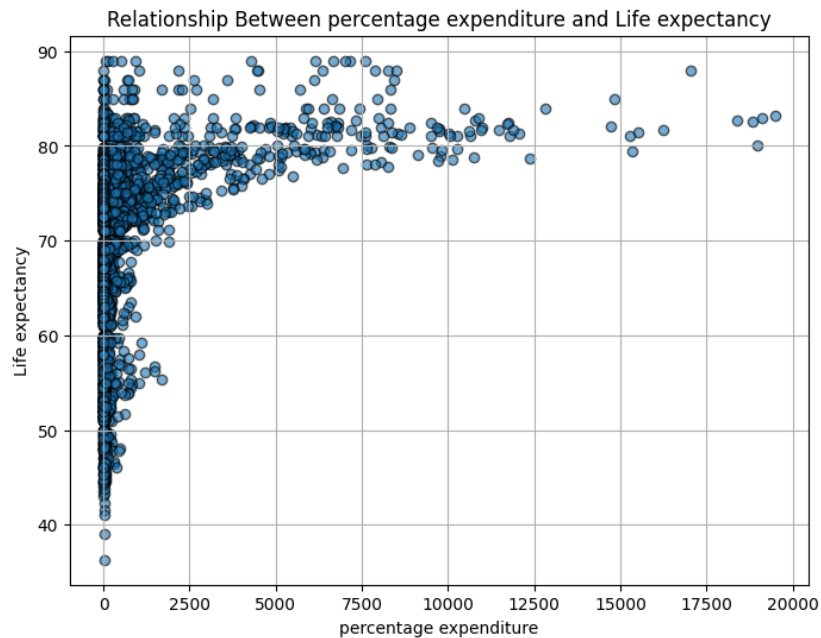
There is a strong negative correlation, with higher adult mortality associated with lower life expectancy.



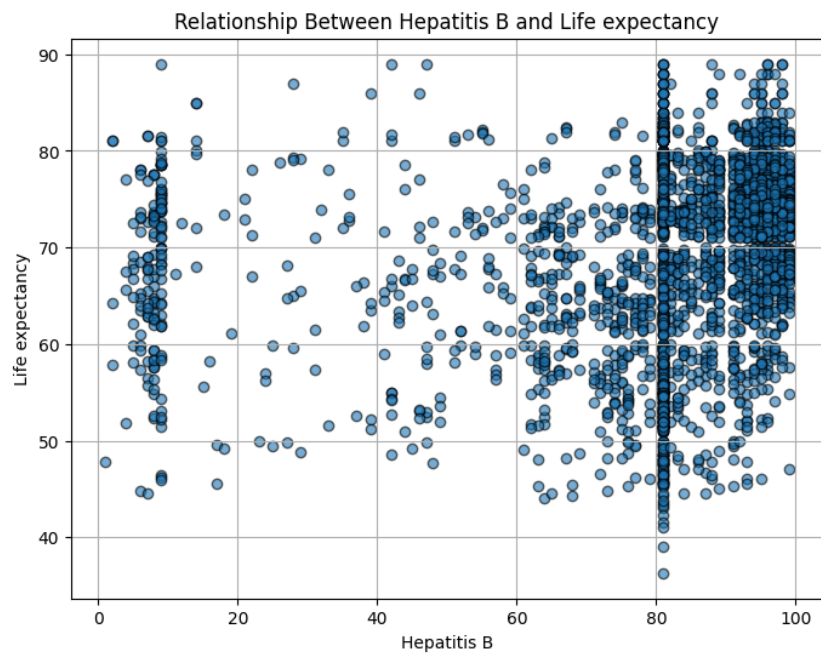
Higher infant death rates are linked to lower life expectancy, indicating the critical impact of early childhood health.



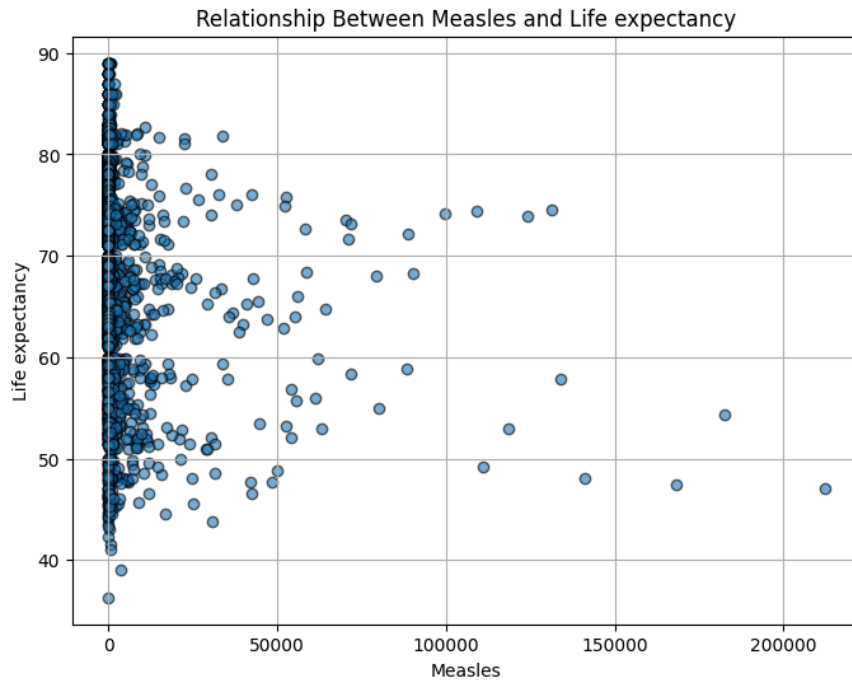
Moderate alcohol consumption appears to correlate positively with life expectancy, though the relationship is non-linear.



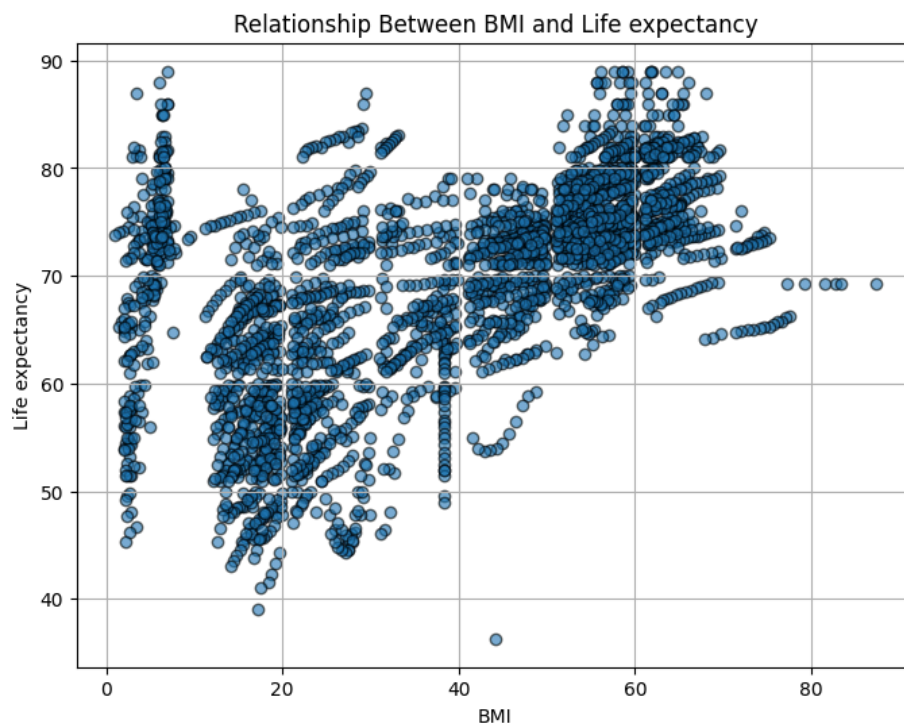
Higher healthcare expenditure as a percentage of GDP is associated with longer life expectancy.



Higher immunization rates for Hepatitis B correspond to increased life expectancy.

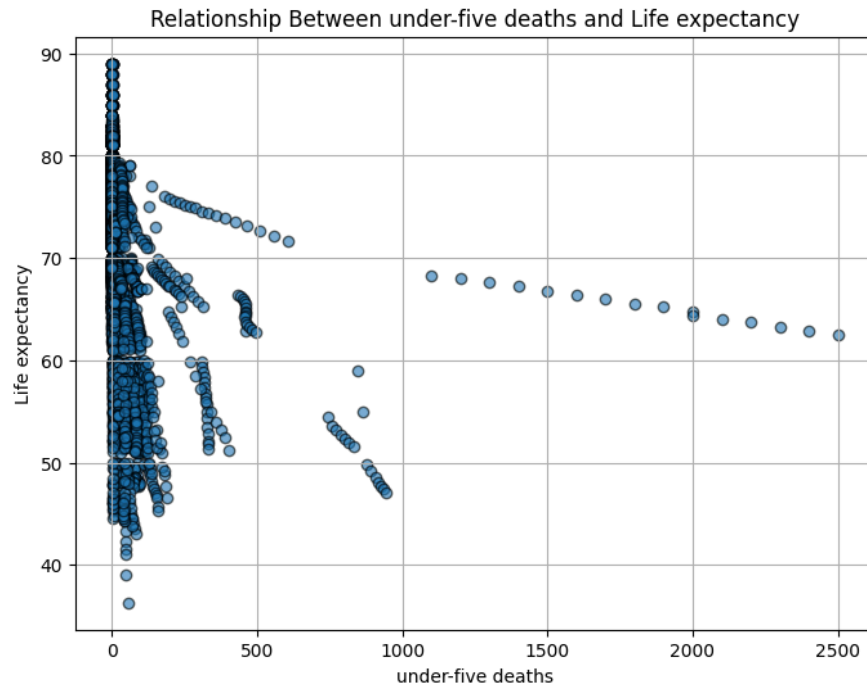


Higher measles cases are linked to lower life expectancy, highlighting the importance of vaccination programs.

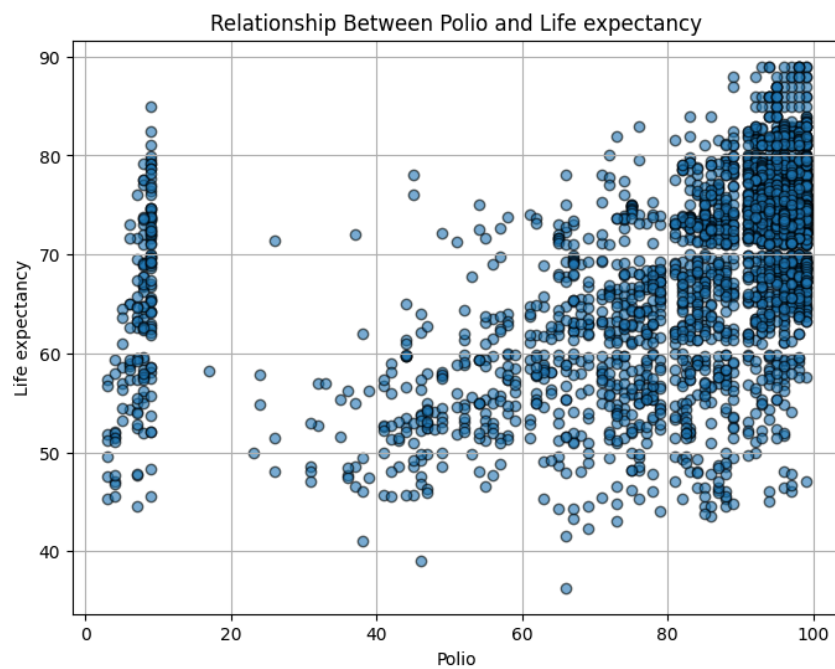


Life expectancy tends to increase with BMI up to a certain threshold, beyond which the relationship weakens.

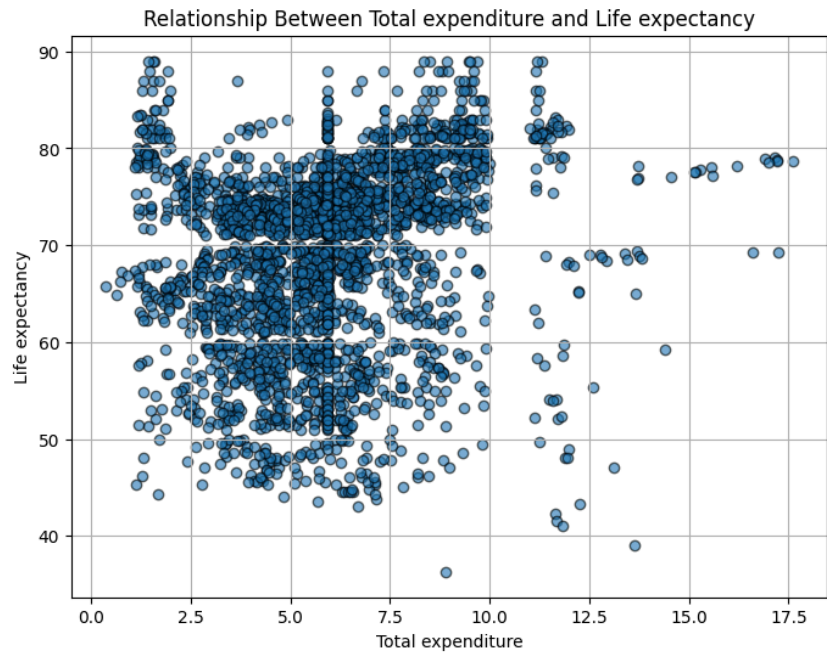




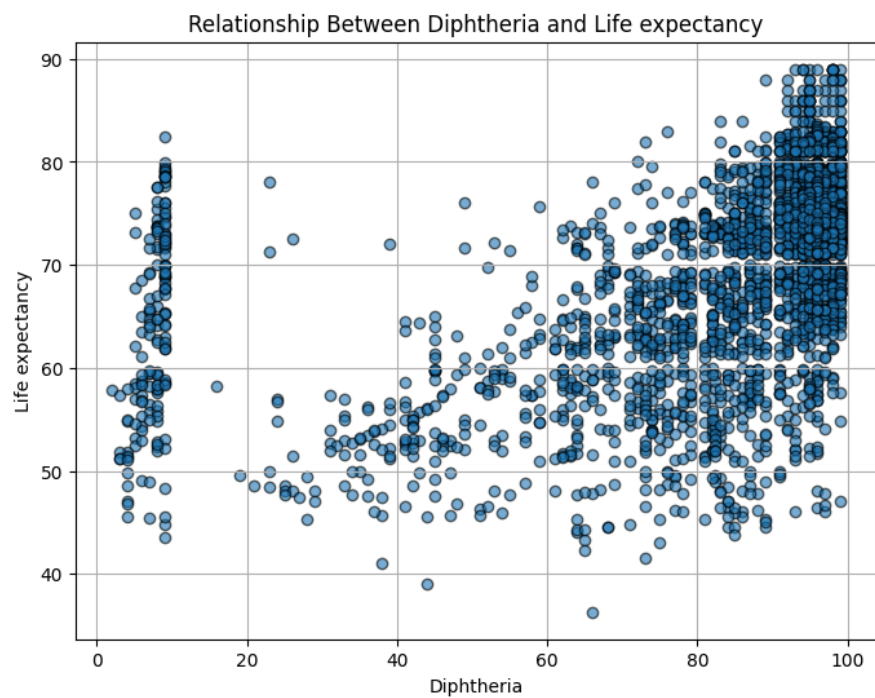
Higher under-five mortality rates significantly reduce life expectancy, underscoring the importance of child health initiatives.



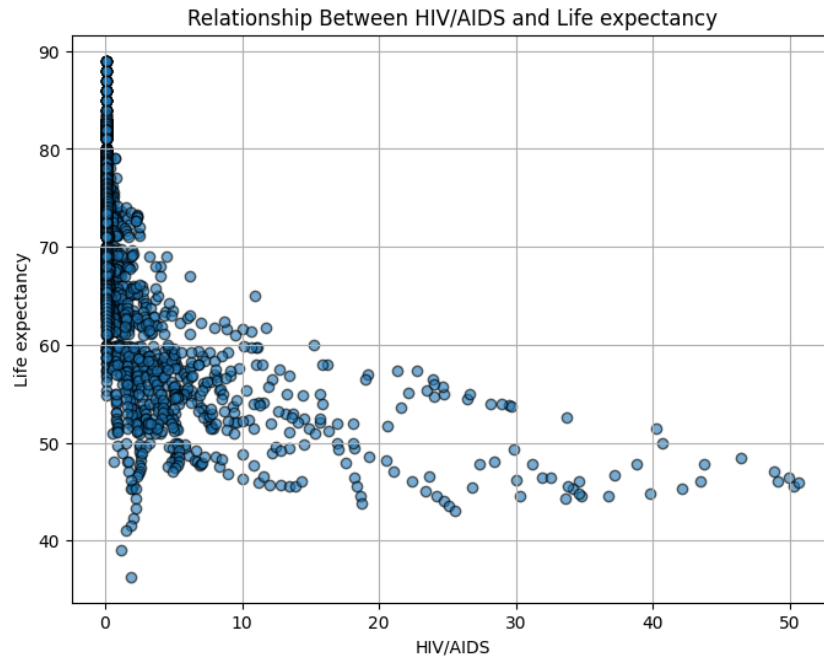
Higher polio immunization rates are positively correlated with increased life expectancy.



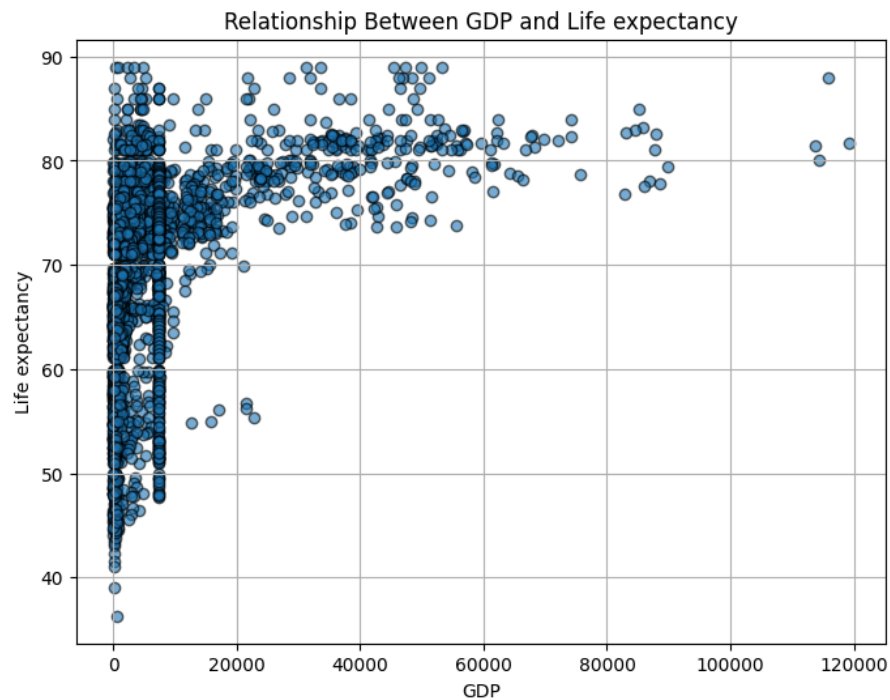
Greater health expenditure shows a weak positive association with life expectancy.



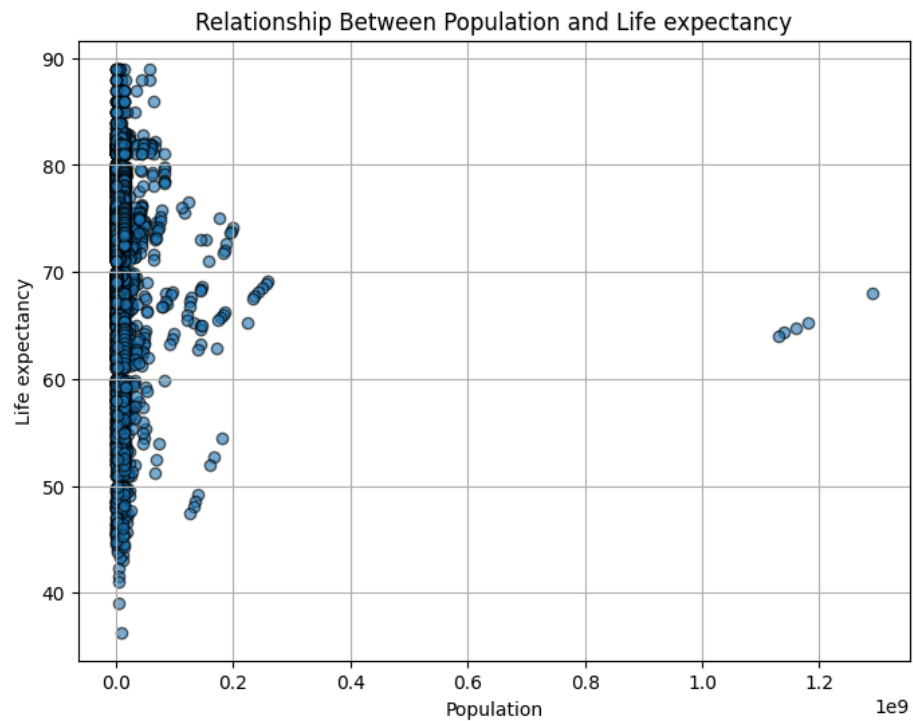
Higher diphtheria immunization rates are strongly linked to longer life expectancy.



Higher HIV/AIDS death rates are negatively correlated with life expectancy.



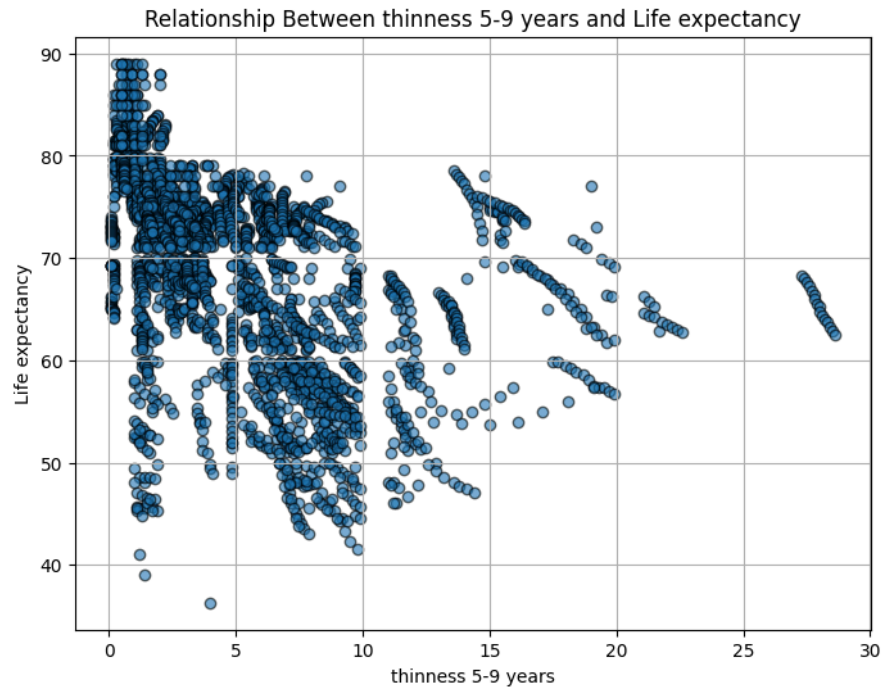
Countries with higher GDP per capita generally have longer life expectancy, though with diminishing returns.



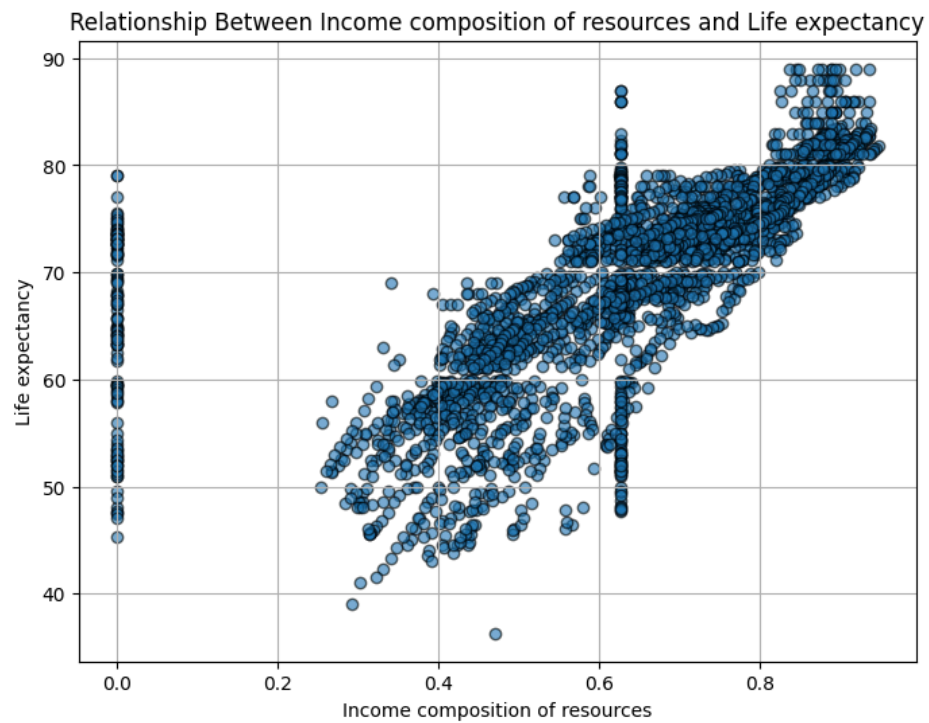
No clear relationship is observed between population size and life expectancy.



Higher rates of thinness in individuals aged 1-19 years are associated with lower life expectancy.



Increased thinness in children aged 5-9 years is linked to reduced life expectancy.

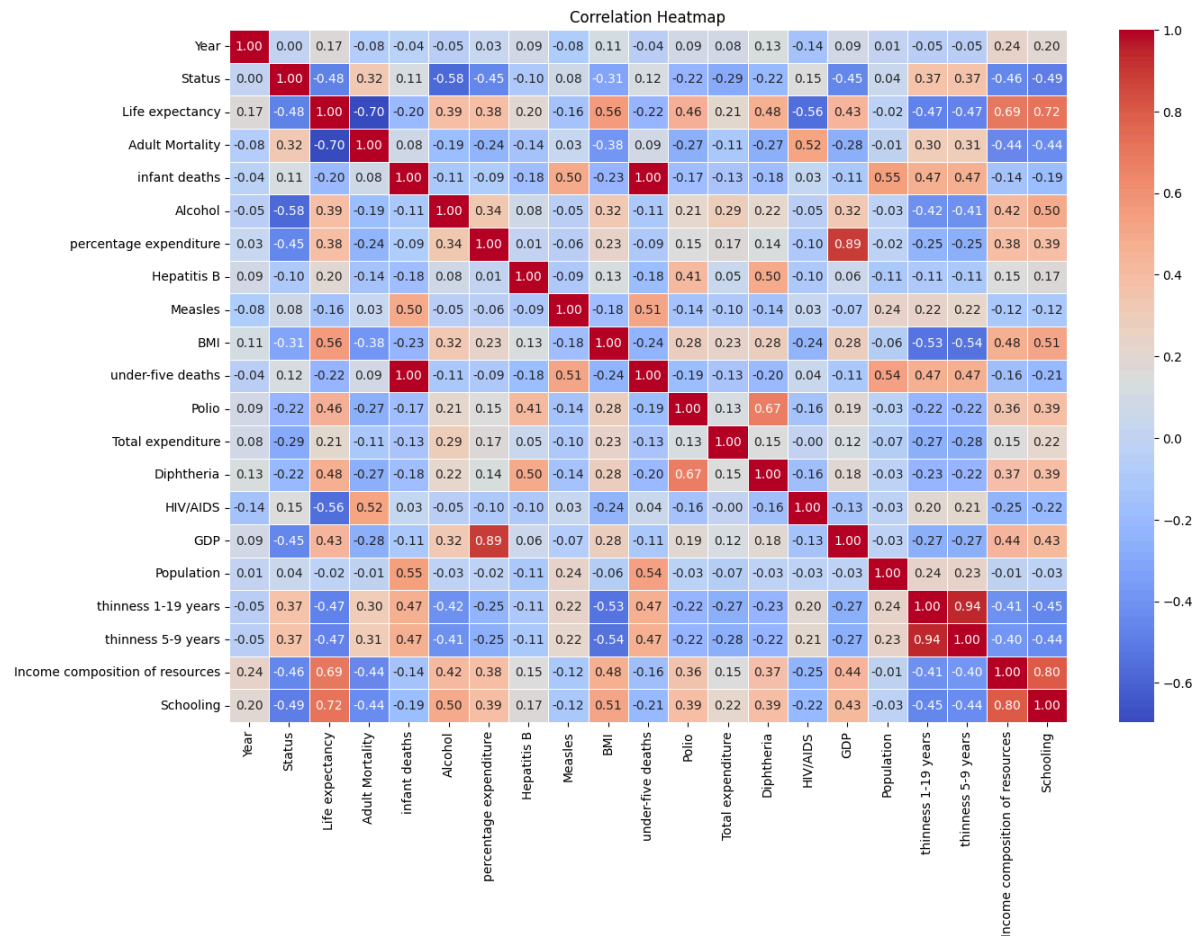


Better income composition of resources correlates strongly with higher life expectancy.



Increased years of schooling are positively correlated with life expectancy, emphasizing the role of education in health outcomes.

## Correlation Heatmap:



The correlation heatmap was computed to understand the relationship between the target variable and the other variables.

### Primary Predictors :

Schooling, Income Composition, Adult Mortality, and Under-Five Deaths are the strongest drivers of life expectancy.

### Secondary Factors :

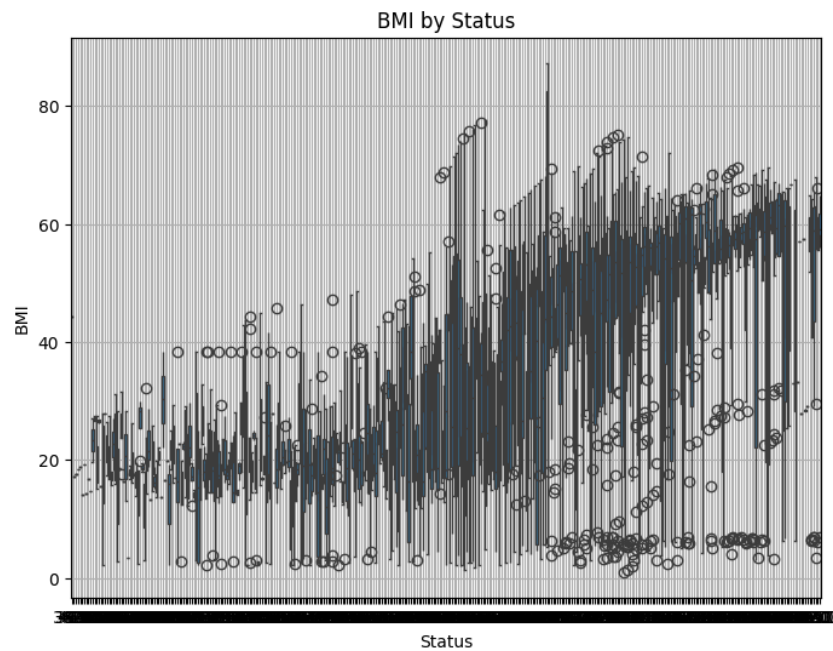
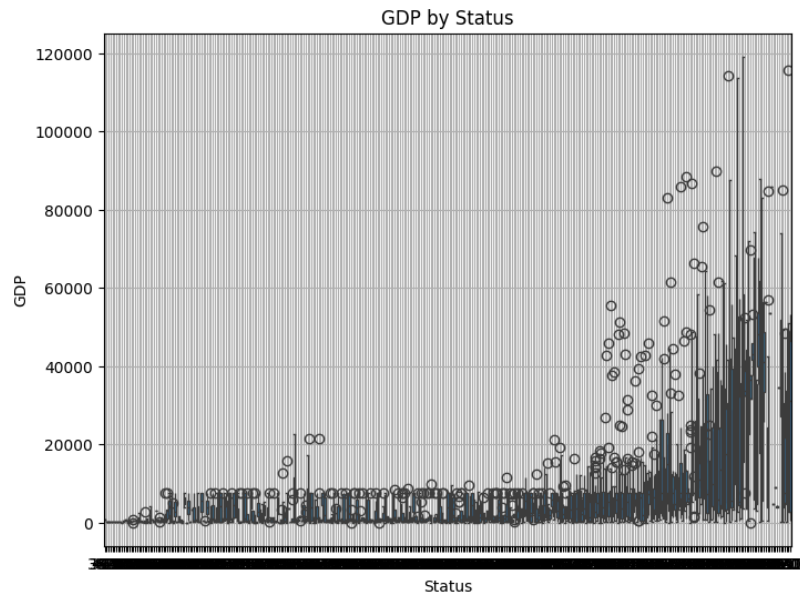
GDP and BMI also contribute but with lesser impact.

### Heatmap Observations:

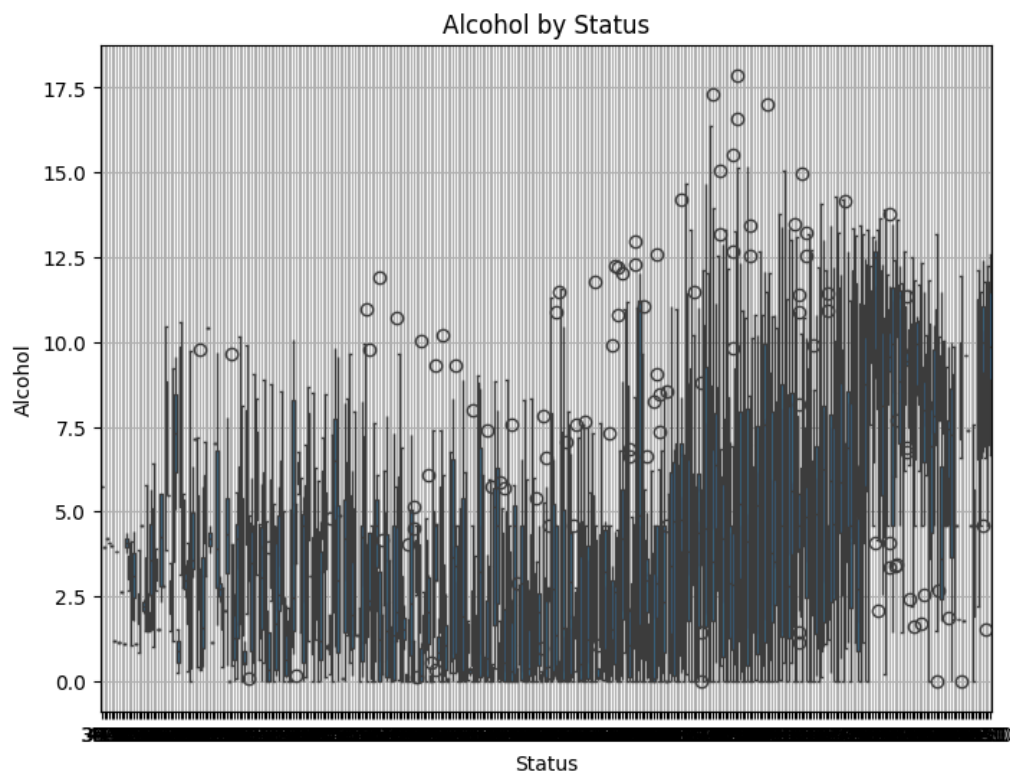
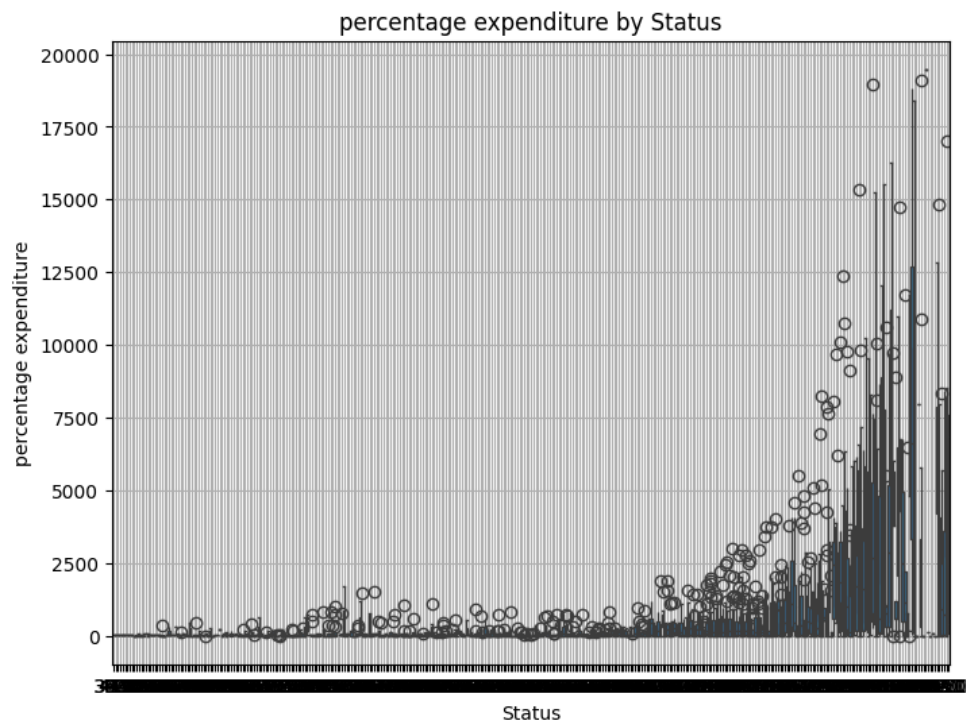
Multicollinearity exists among some predictors (e.g., Income Composition of Resources and Schooling), which may affect regression models.

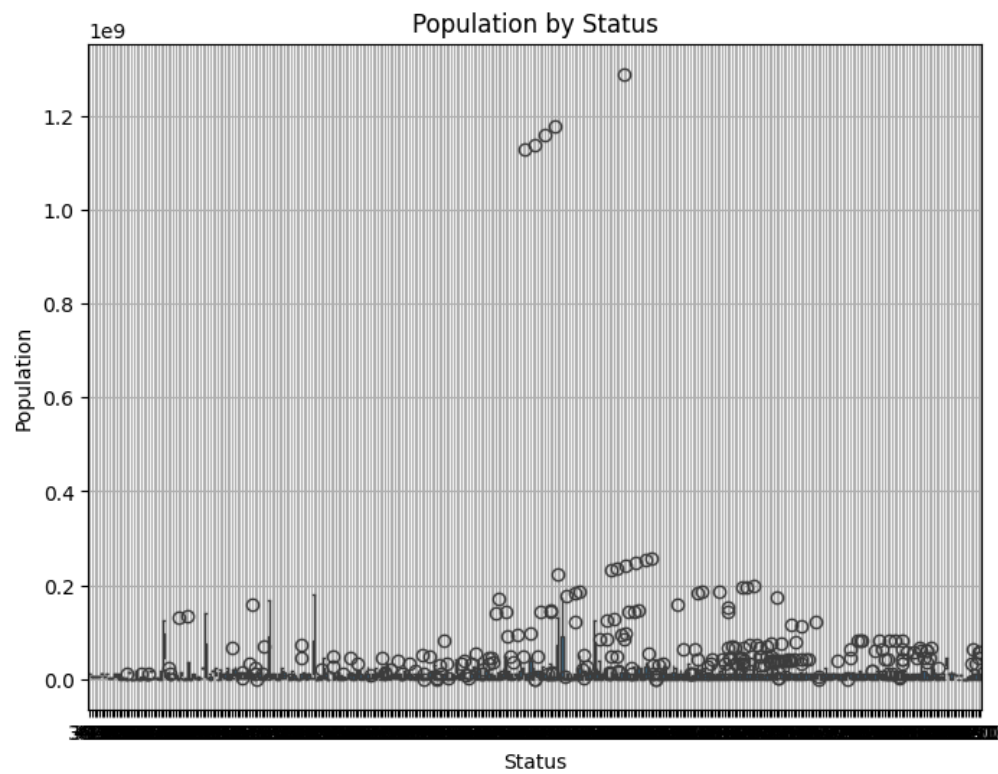
Features with strong positive and negative correlations with life expectancy should be prioritized for modeling.

Further, analysis was conducted to explore the relationship between strong predictors and the categorical variable status. Following are the findings:

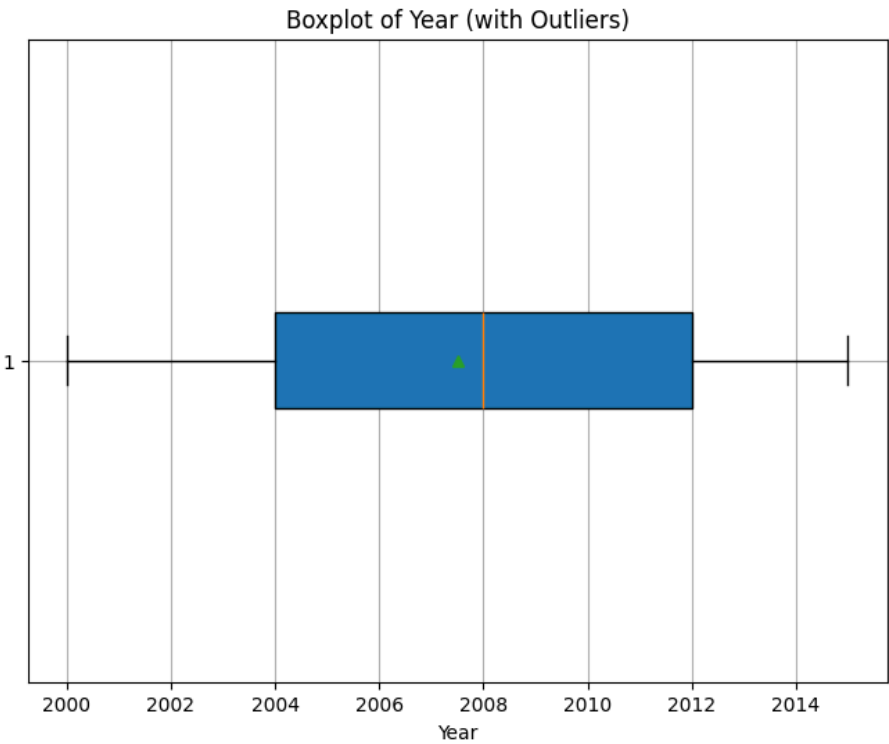


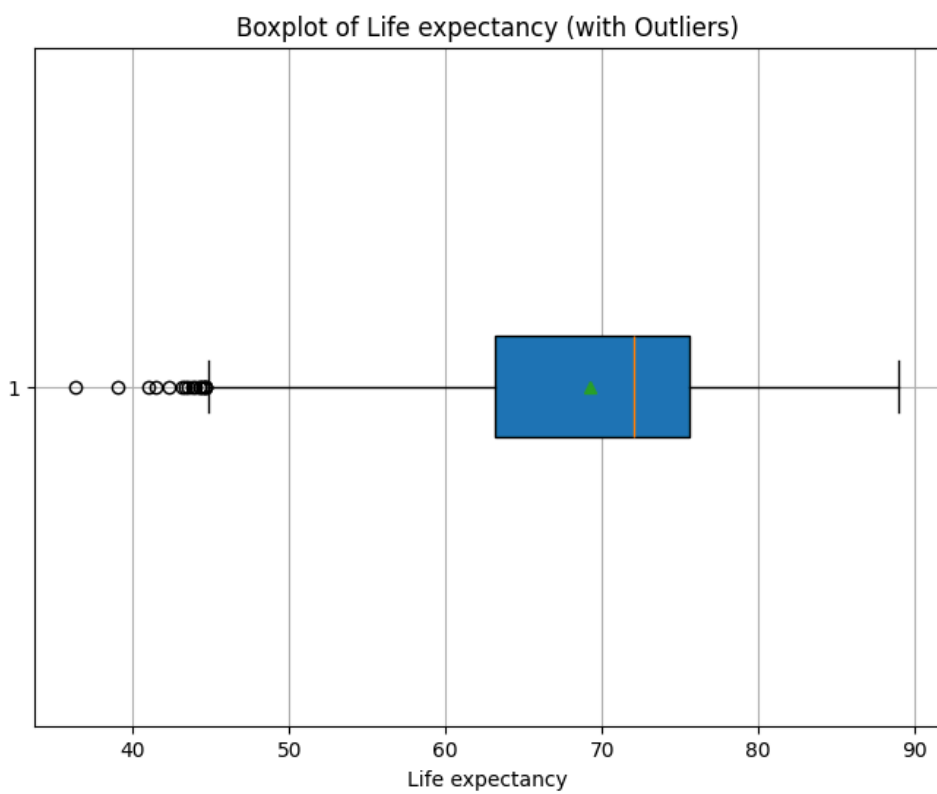
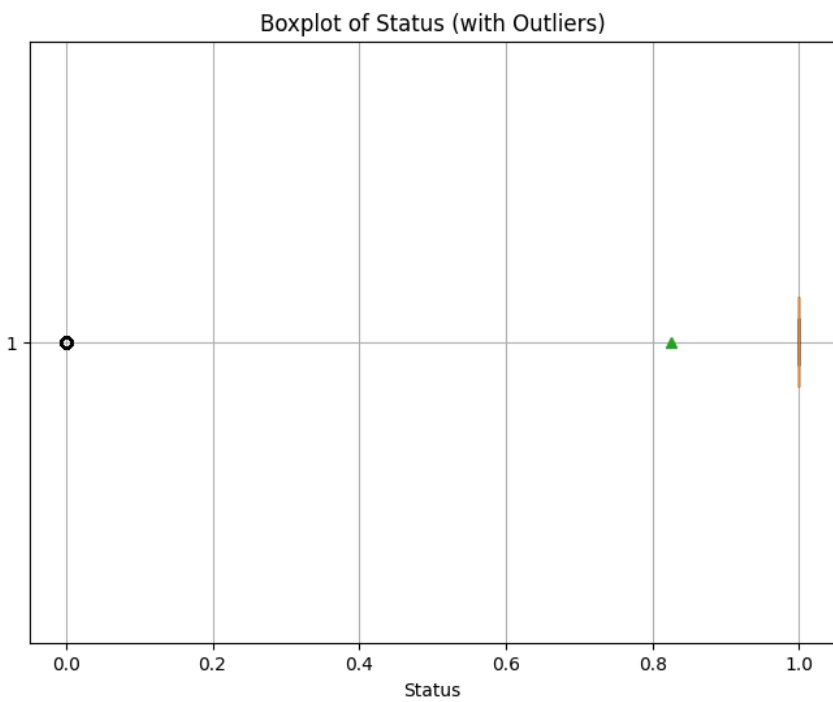


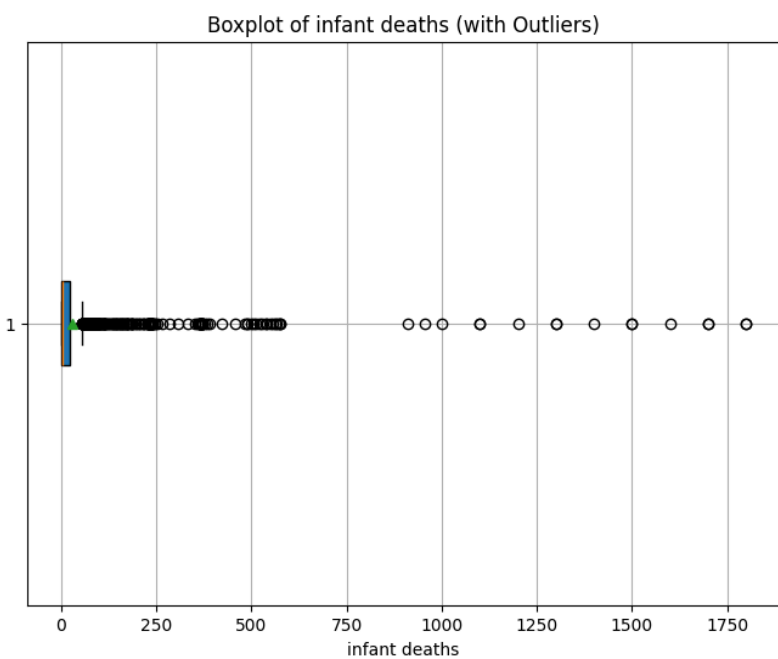
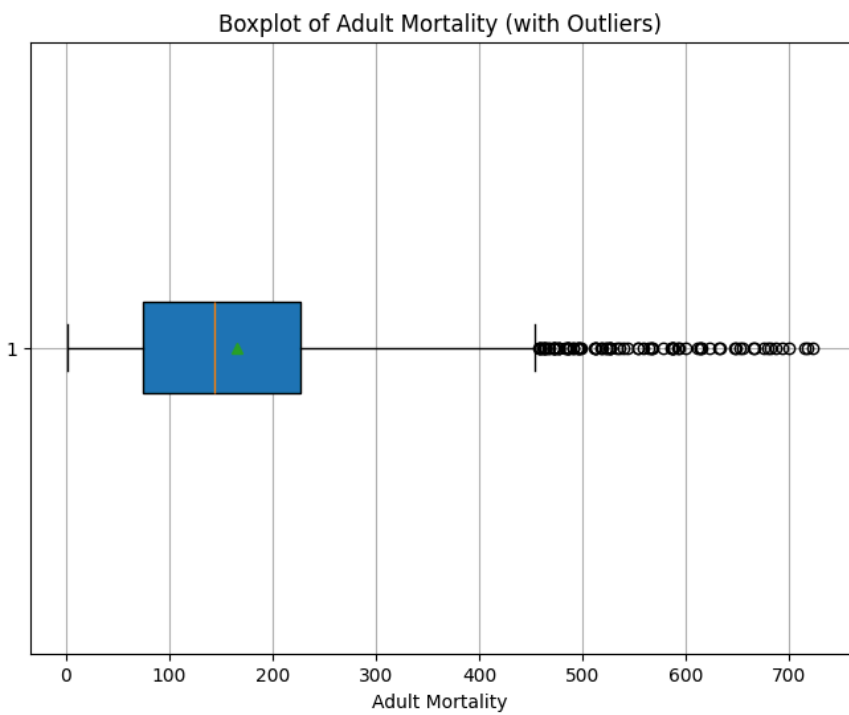


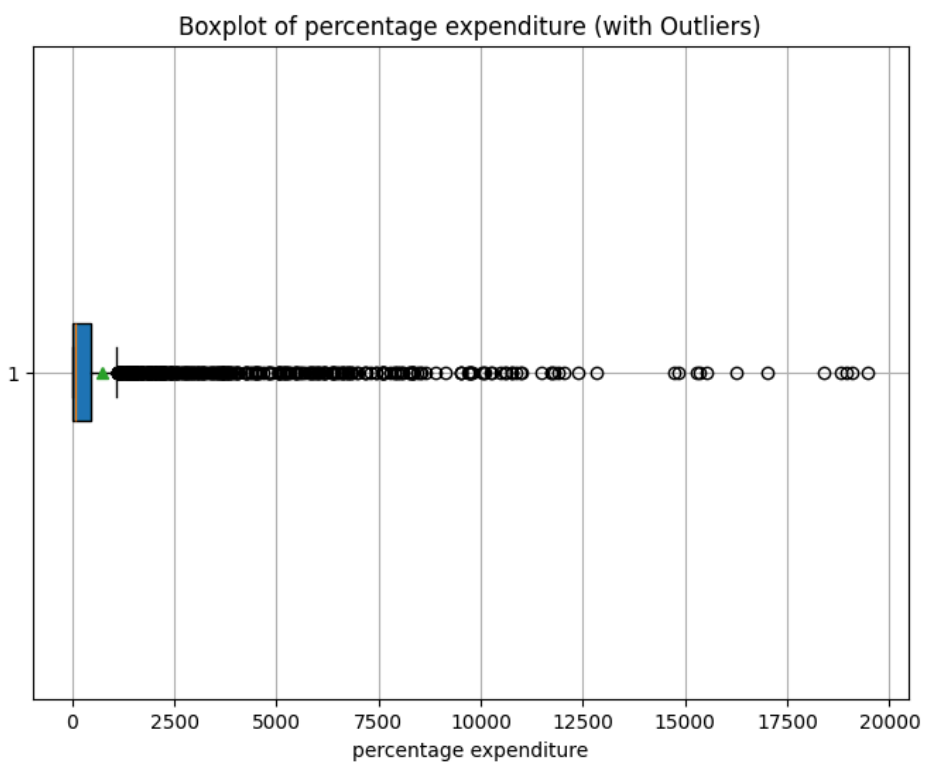
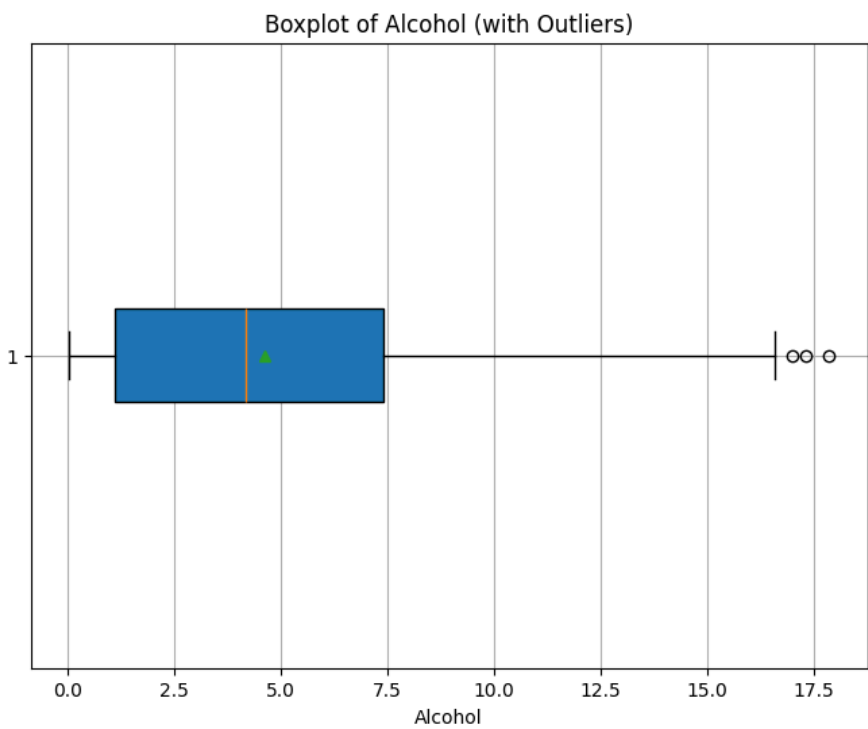


Boxplots of numeric variables with outliers:

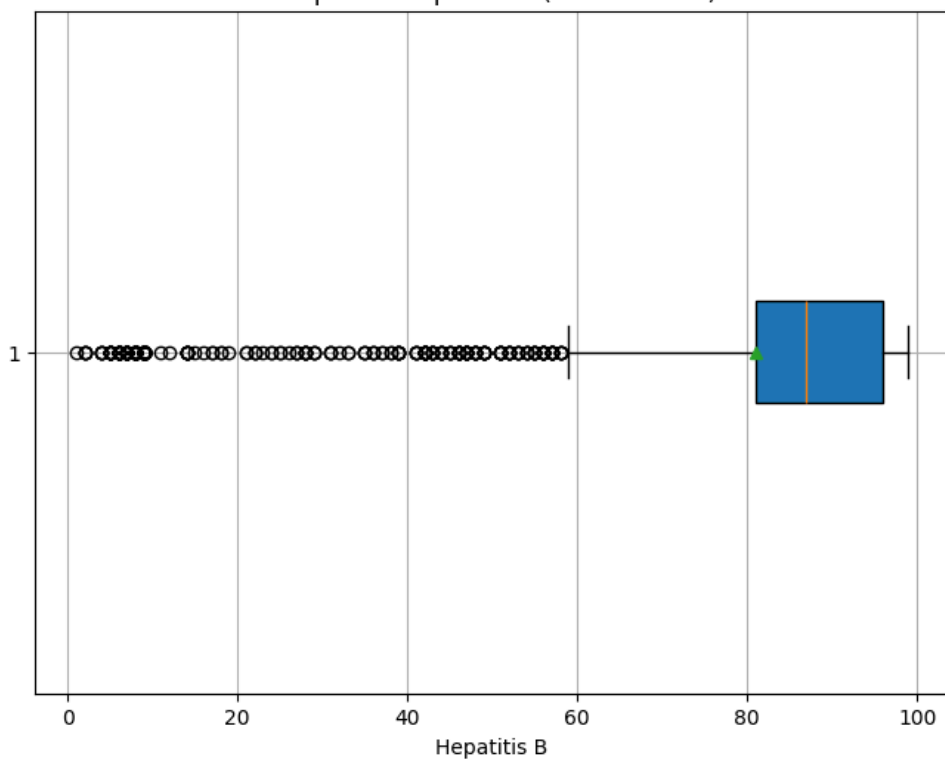




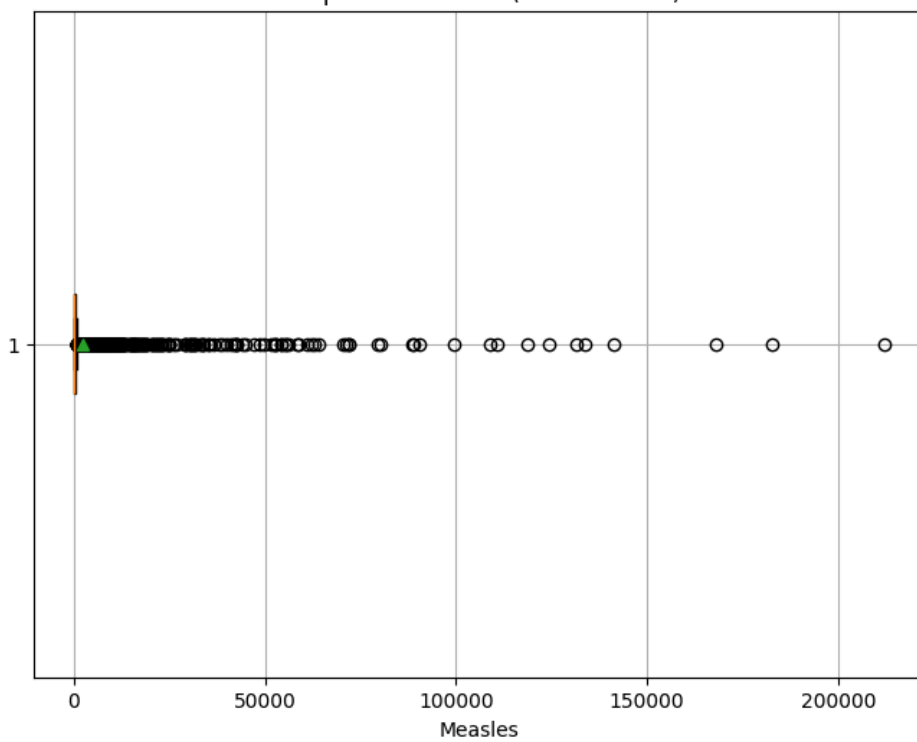


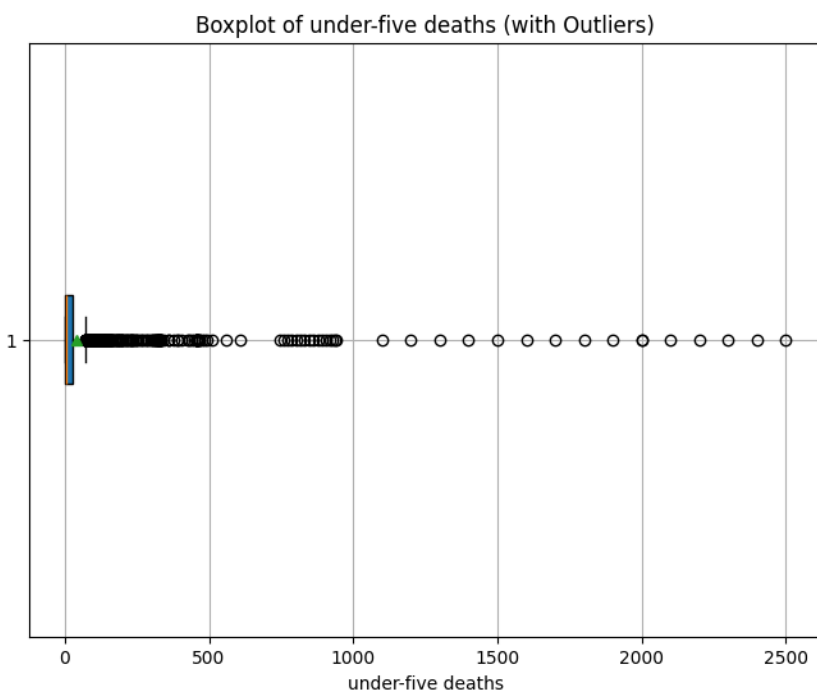
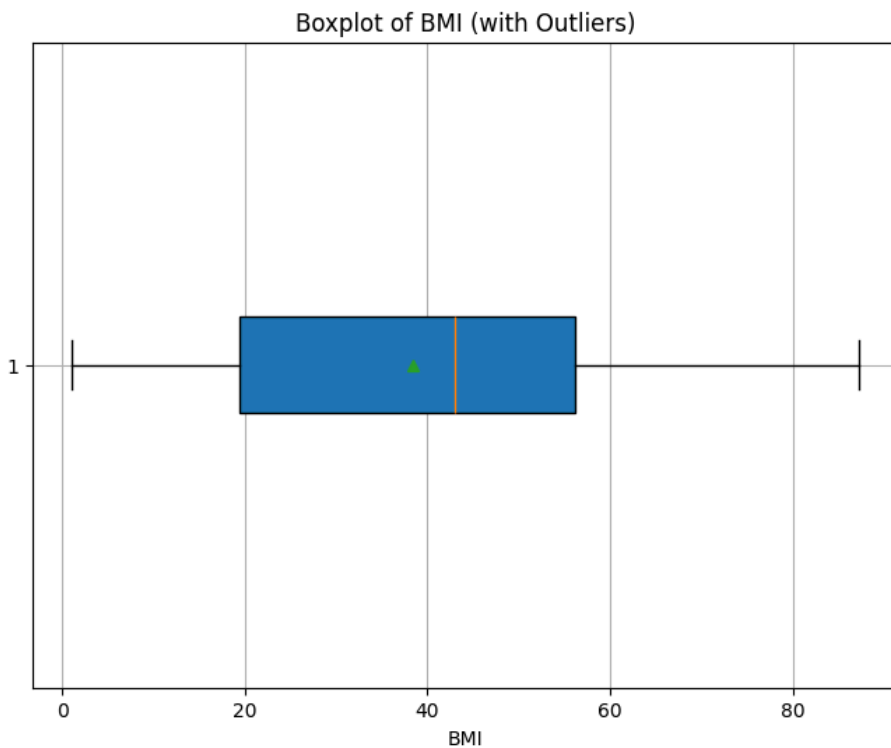


Boxplot of Hepatitis B (with Outliers)

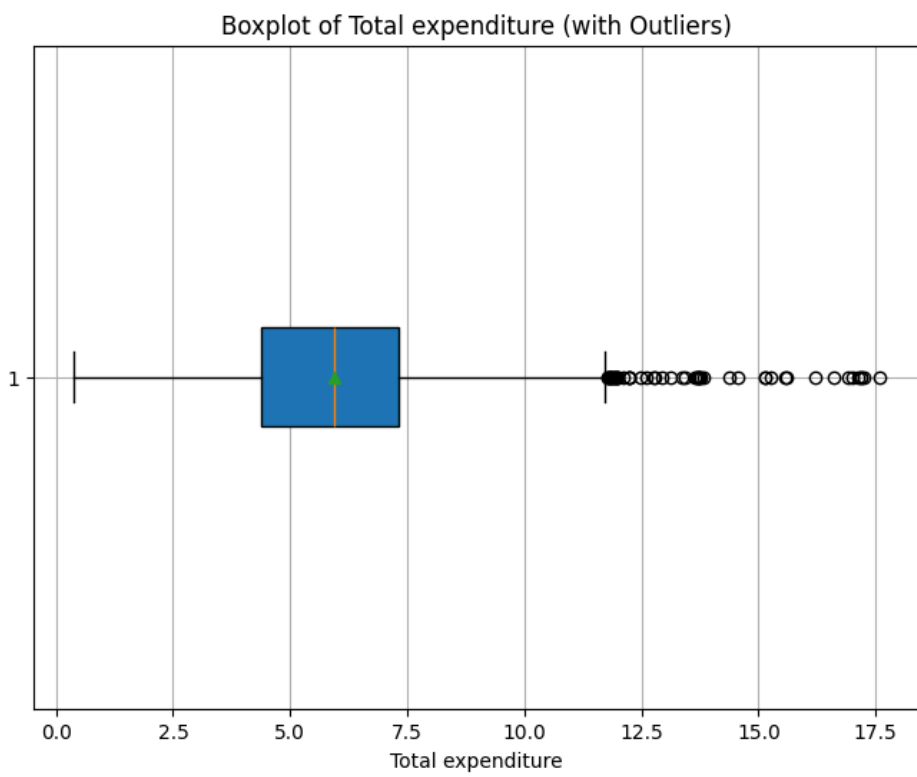
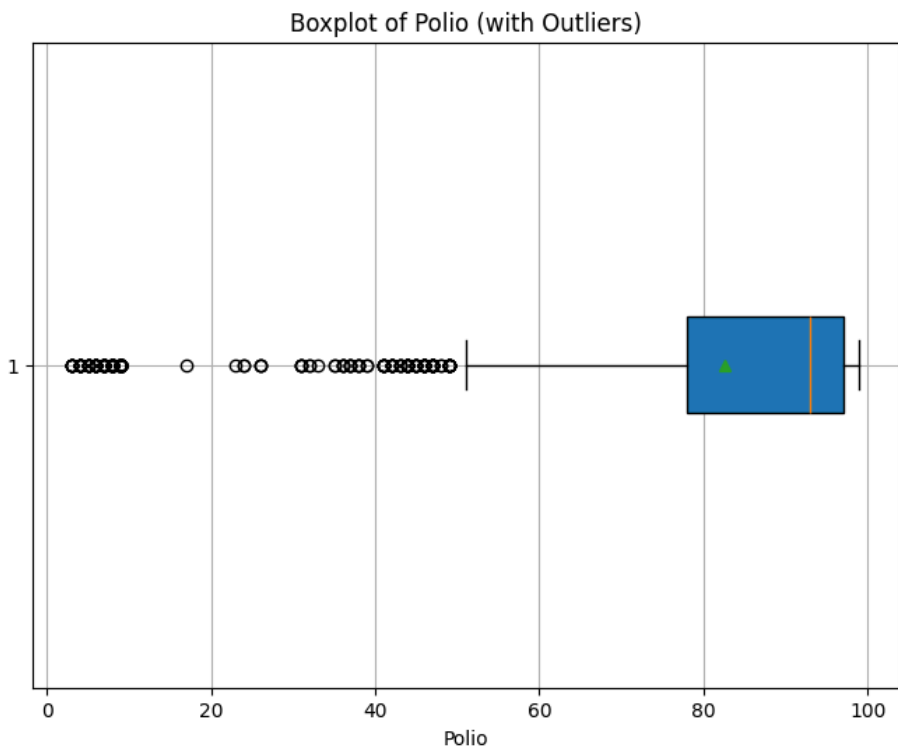


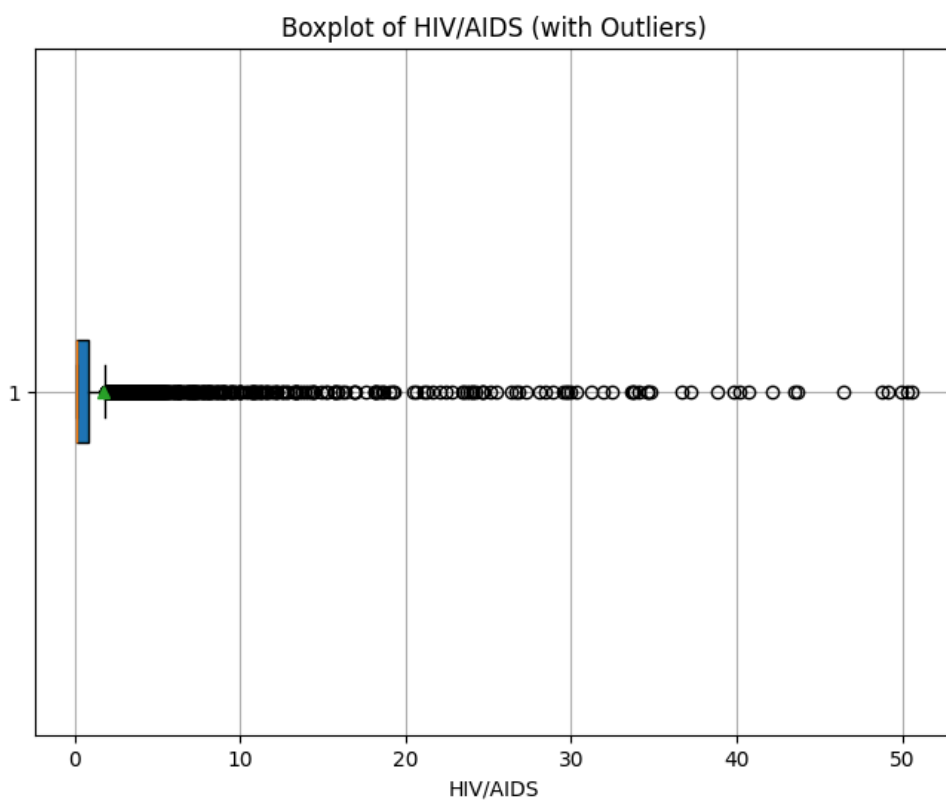
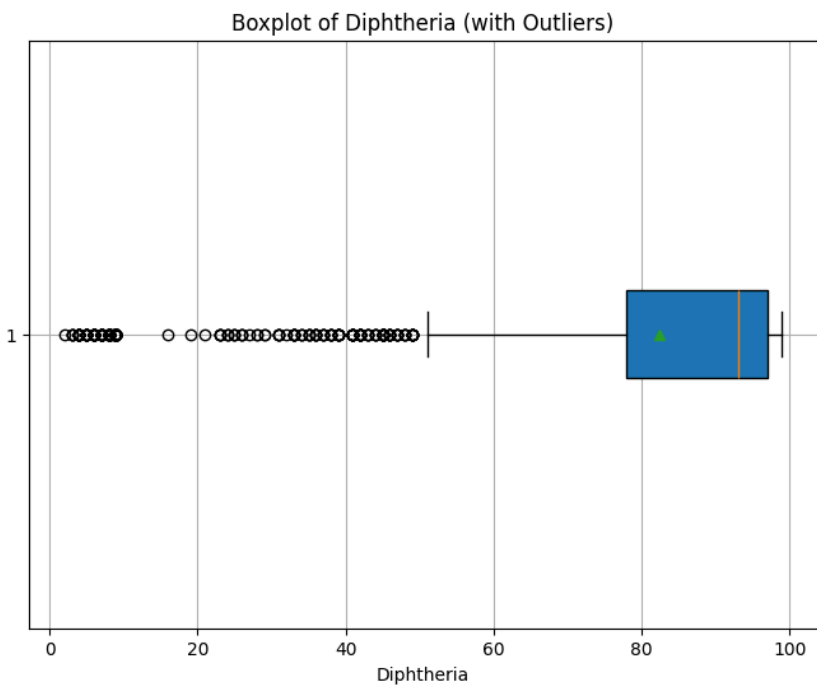
Boxplot of Measles (with Outliers)

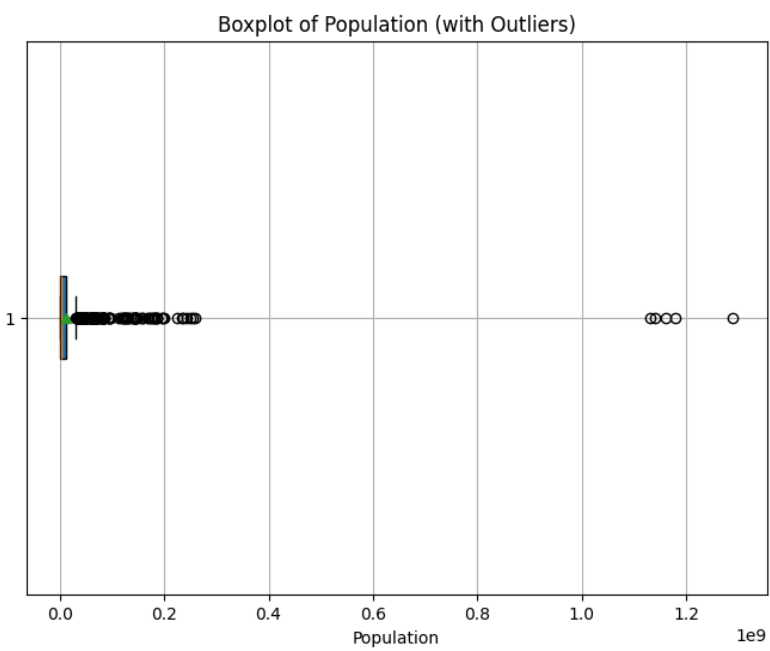
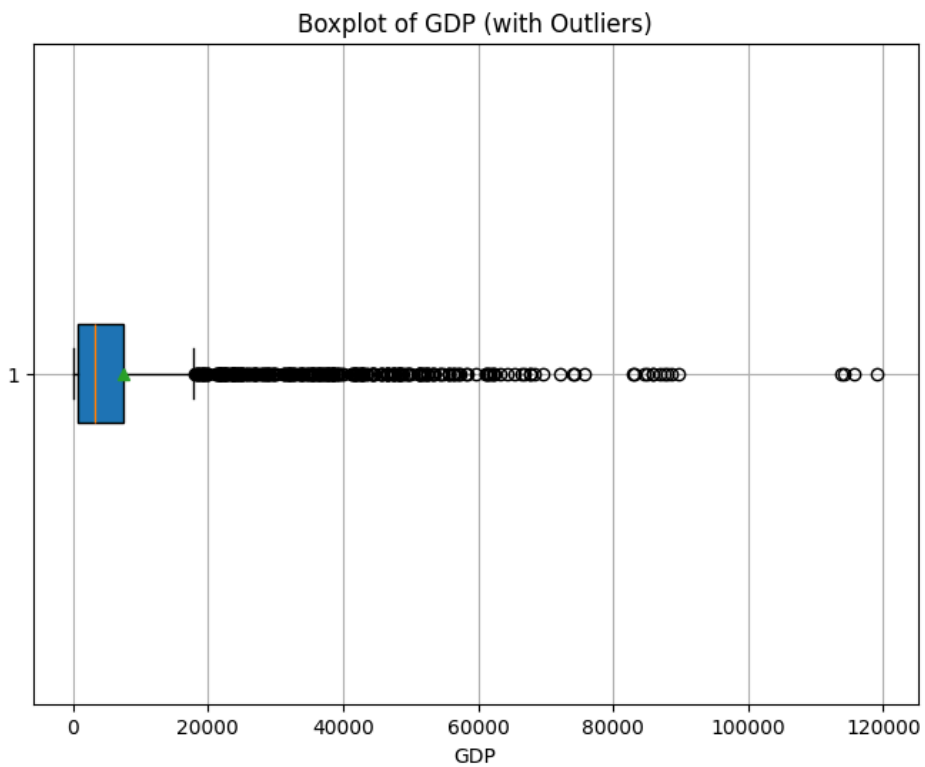




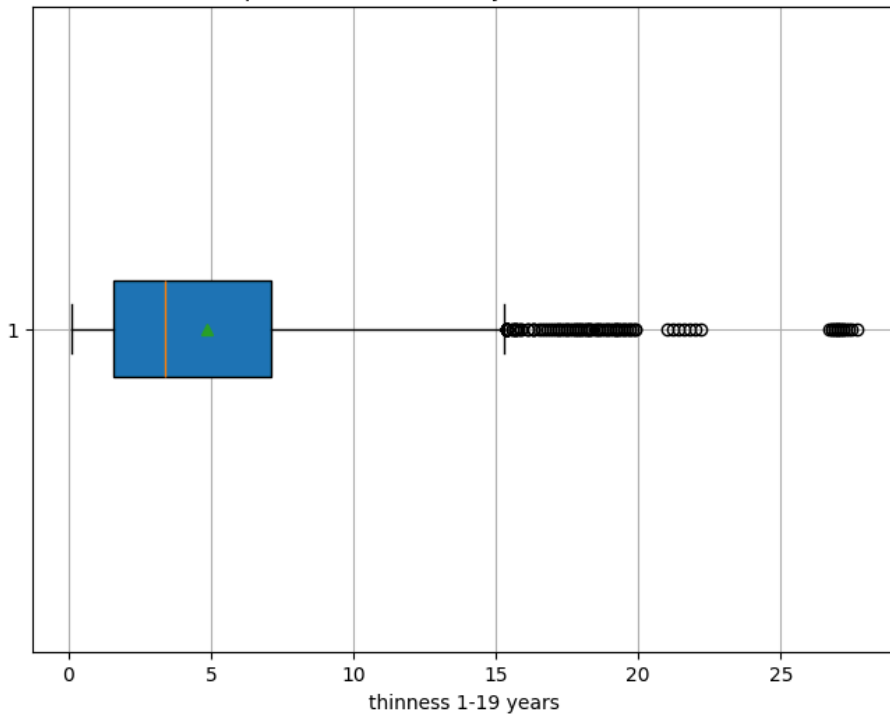




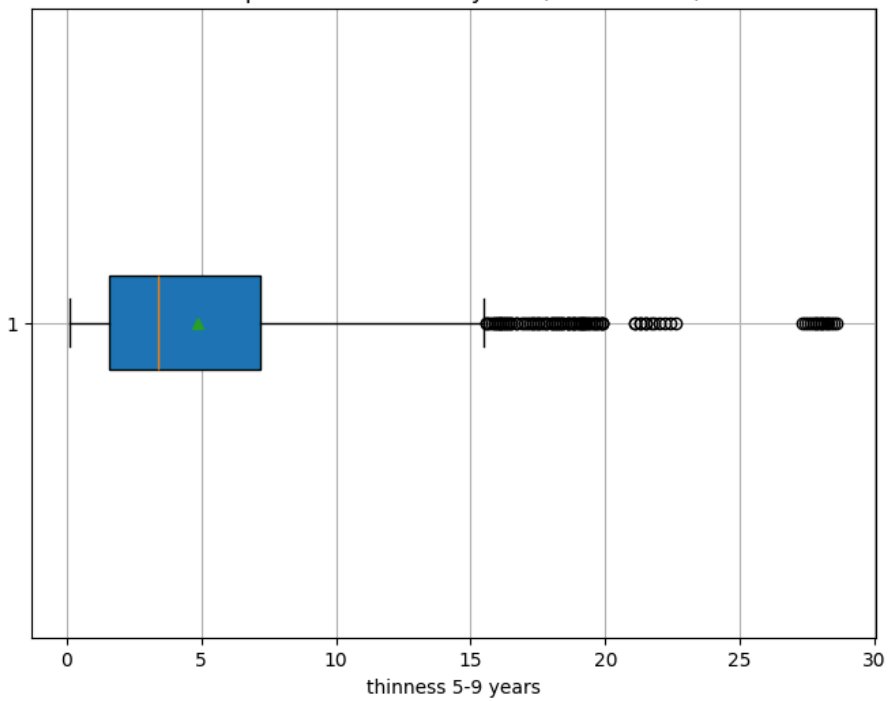


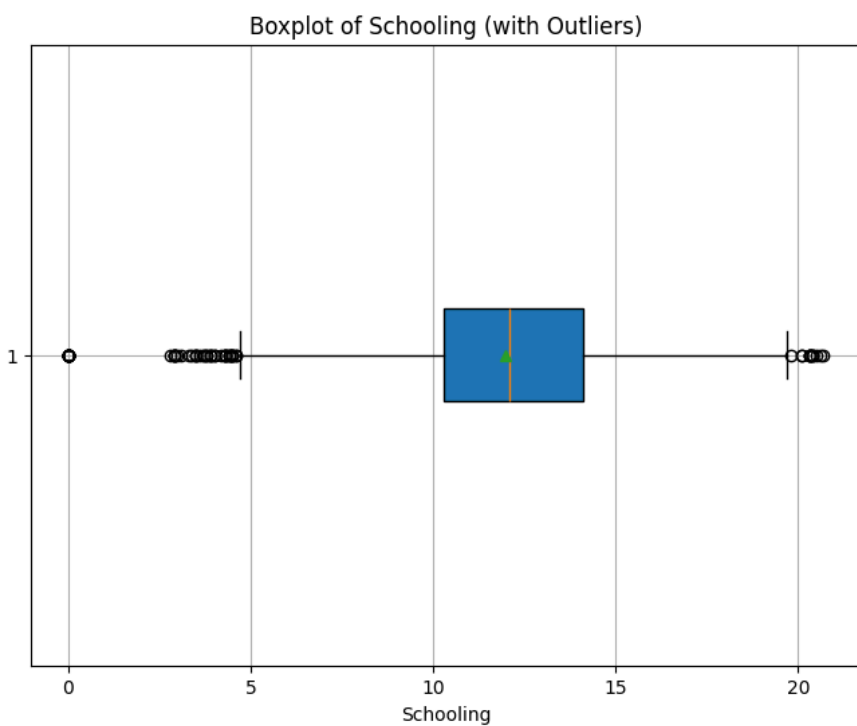
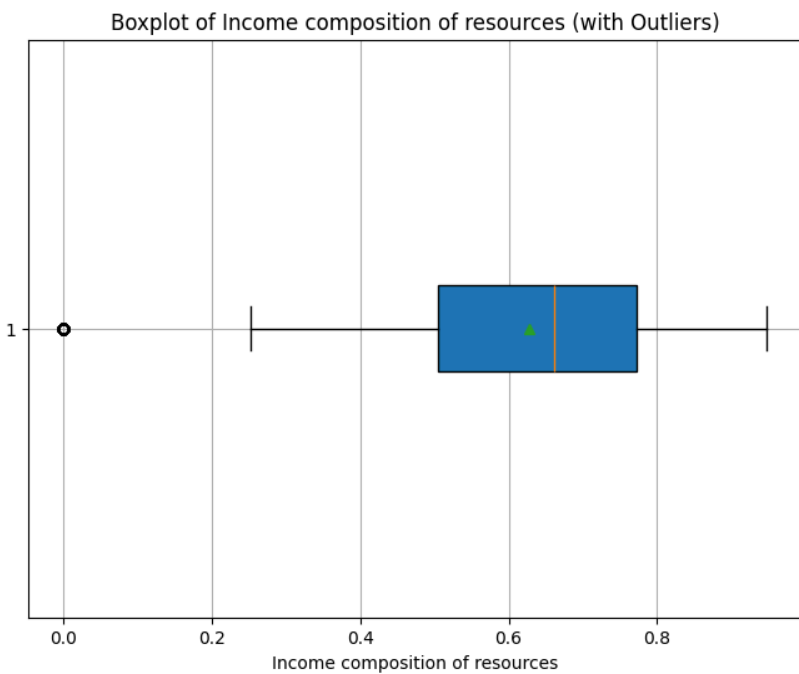


Boxplot of thinness 1-19 years (with Outliers)



Boxplot of thinness 5-9 years (with Outliers)



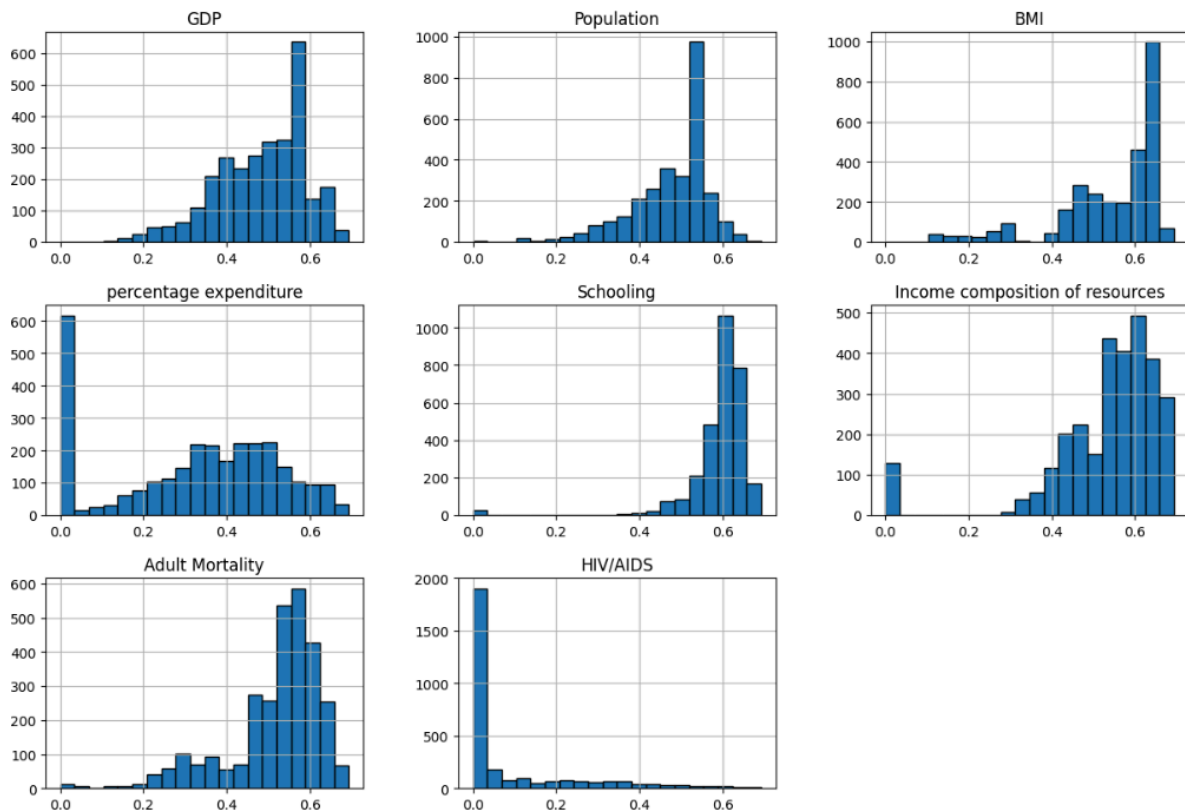


## Handled Outliers:

Outliers were detected and those which were strong predictors were handled using log transformation.



Histograms of Log-Transformed Features



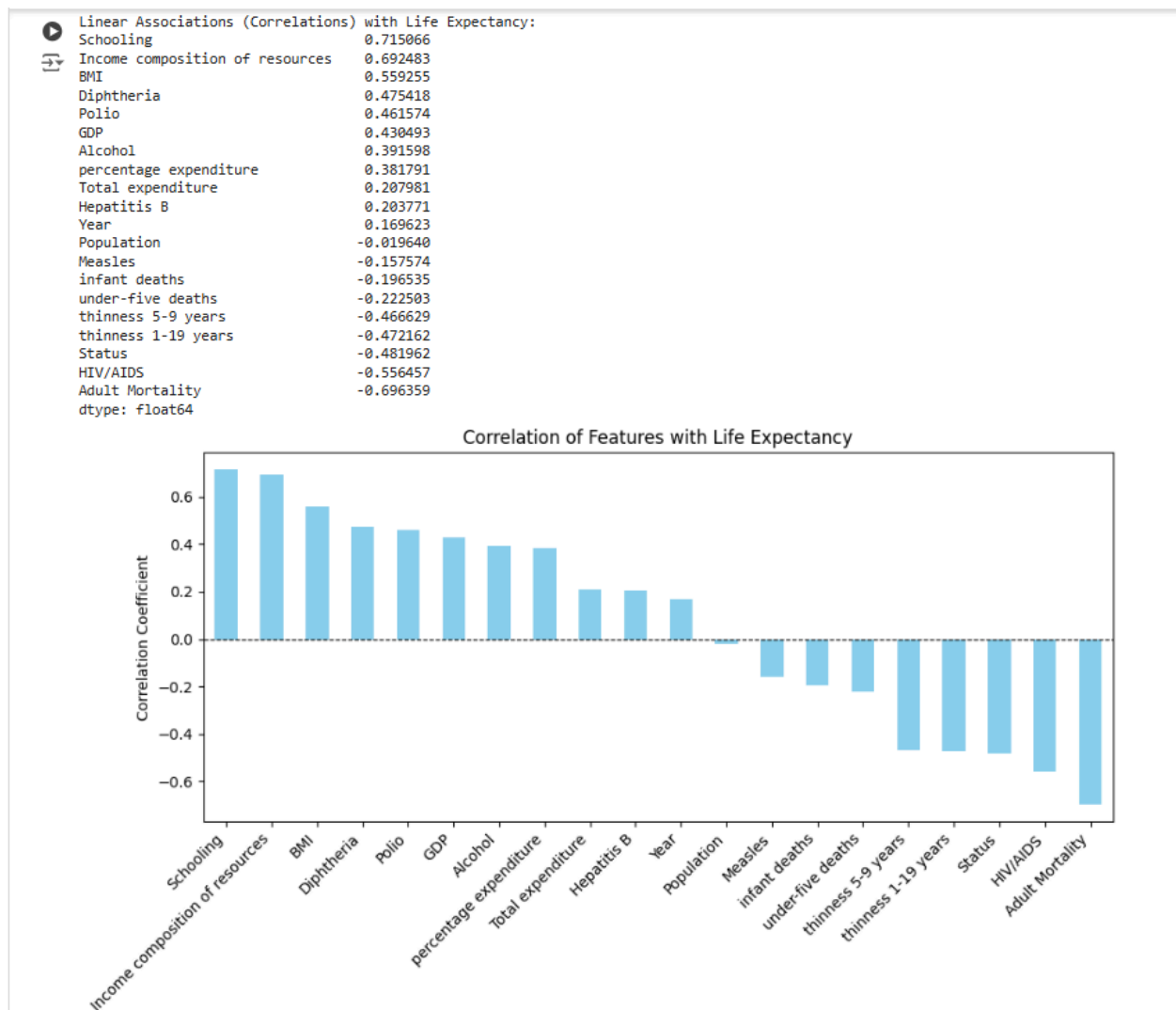
## Linear Association

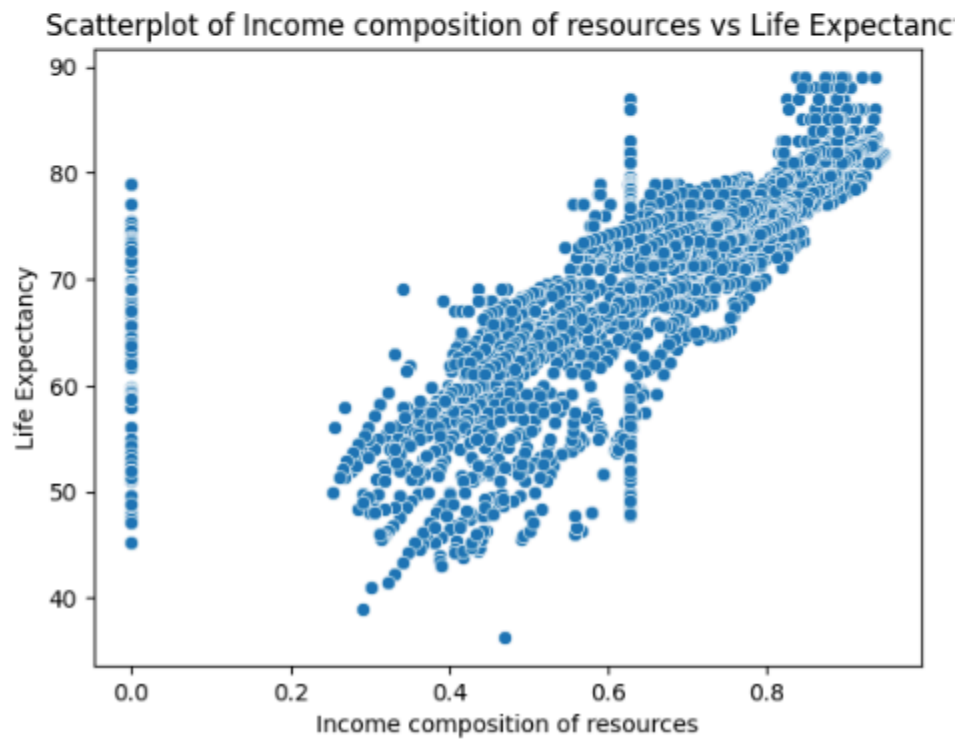
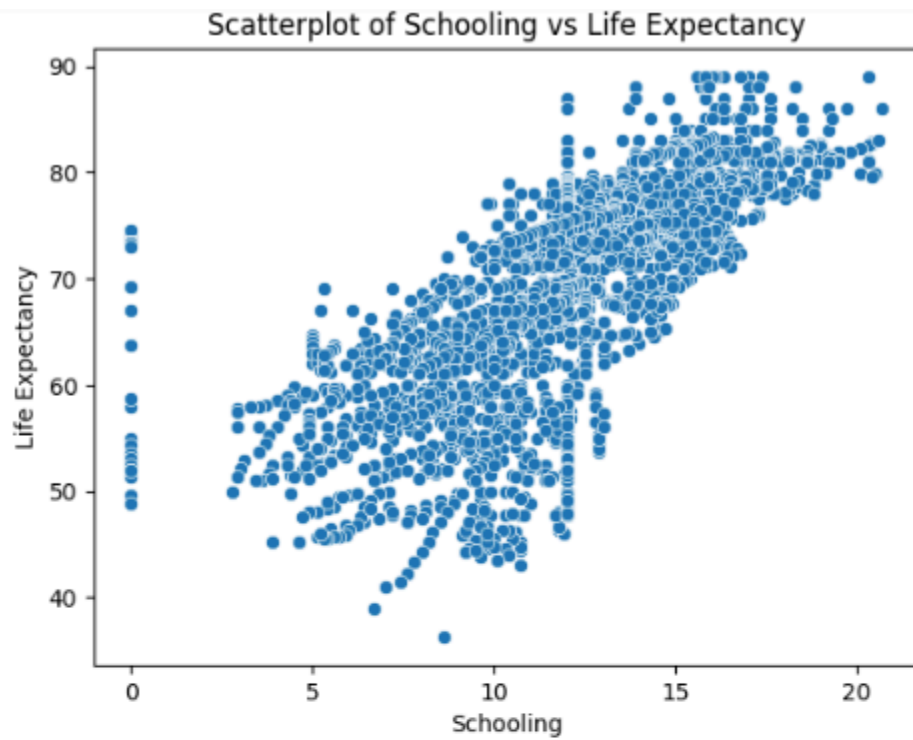
Linear relationships are described by the relationship between two variables where a change in one variable is directly proportional to a change in the other. In other words, if one variable increases (decreases), the other variable tends to increase (decrease) in a manner that is consistent. The relationship can be positive, negative, or neutral.

Positive Relationship: To explain, if there is a positive relationship, then with an increase in one variable, there will be an increase in the value of the other variable as well.

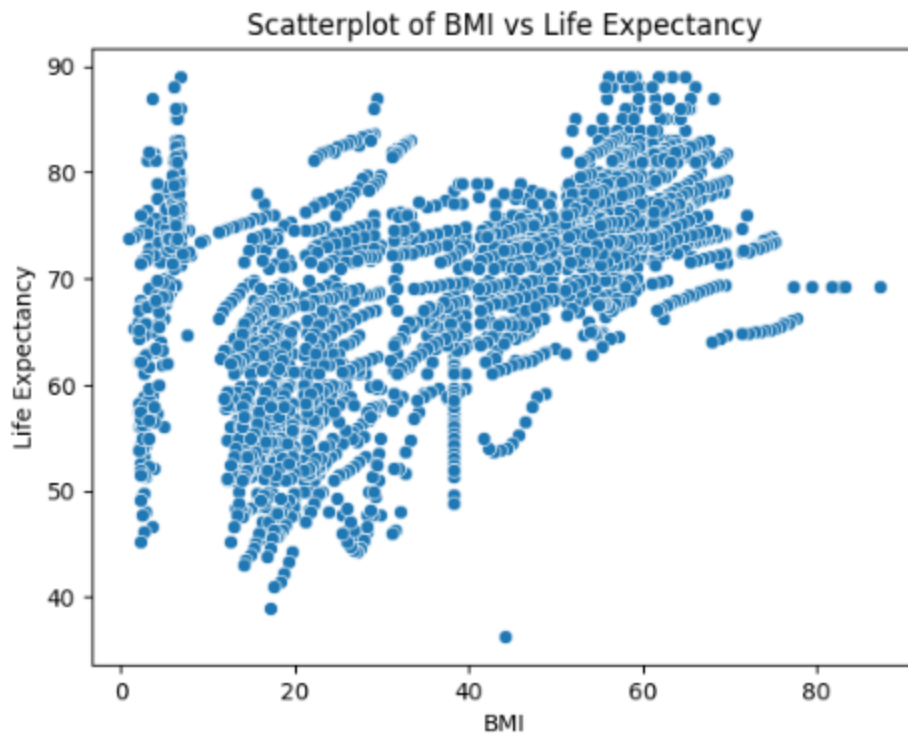
Negative Linear Association: As one variable increases, the other variable decreases.

Example: Higher pollution levels may be associated with lower air quality.









- **Strong Positive Correlations** with Life Expectancy

- Schooling (0.715):
  - Education level is the most strongly positively correlated variable with life expectancy. Higher schooling levels are associated with longer life expectancy.
- Income Composition of Resources (0.692):
  - A higher income composition index, indicating better access to resources, strongly correlates with higher life expectancy.
- BMI (0.559):
  - A healthier body mass index (within reasonable ranges) is positively associated with life expectancy.
- Vaccination Rates (Diphtheria: 0.475, Polio: 0.462):
  - Vaccination coverage shows a meaningful positive relationship with life expectancy, indicating the importance of immunization in public health.

- **Strong Negative Correlations** with Life Expectancy

- Adult Mortality (-0.696):
  - The strongest negative correlation. Higher adult mortality rates directly reduce life expectancy.
- HIV/AIDS (-0.556):
  - A significant negative relationship highlights the impact of diseases on life expectancy.

- Thinness in Children (10-19 years: -0.472, 5-9 years: -0.466):
  - Malnutrition and thinness among children are strongly linked to reduced life expectancy.
- Status (Developed vs. Developing: -0.482):
  - Countries categorized as "Developing" have significantly lower life expectancy compared to "Developed" countries.

- **Moderate Correlations**

- GDP (0.434):
  - Economic performance moderately correlates with life expectancy. Wealthier nations tend to have higher life expectancies.
- Alcohol Consumption (0.392):
  - Moderate positive correlation. It may reflect economic development or lifestyle changes, but this requires further investigation.
- Percentage Expenditure (0.381):
  - Higher health expenditure relative to GDP moderately correlates with life expectancy, showing the role of healthcare investment.

- **Weak or Negligible Correlations**

- Population (-0.019):
  - Almost no correlation, suggesting population size doesn't directly impact life expectancy.
- Year (0.169):
  - Weak positive correlation; life expectancy has gradually increased over time, but this variable on its own doesn't hold significant predictive power.

## Normalization:

Further the normalisation was done on the dataset using the normalisation function in python. Normalization ensures that all features in the dataset are on a similar scale, preventing variables with larger ranges from dominating those with smaller ranges. This is particularly important for machine learning algorithms like regression, SVM, or KNN, which are sensitive to the magnitude of input features.

## Multicollinearity:

Handled the multicollinearity using Variable Inflation Factor (VIF)

	Feature	VIF
0	Year	1.151309
1	Status	1.900893
2	Life expectancy	5.507714
3	Adult Mortality	1.964653
4	infant deaths	184.487789
5	Alcohol	1.863300
6	percentage expenditure	5.003610
7	Hepatitis B	1.406419
8	Measles	1.385227
9	BMI	1.696303
10	under-five deaths	184.035819
11	Polio	1.975965
12	Total expenditure	1.206220
13	Diphtheria	2.272904
14	HIV/AIDS	1.895758
15	GDP	5.143887
16	Population	1.488289
17	thinness 1-19 years	8.782087
18	thinness 5-9 years	8.870742
19	Income composition of resources	2.589693
20	Schooling	3.204317

Further, the dataset was split into train and test dataset and models were applied. Cross validation approach was used

Cross-validation is a resampling method that assesses model performance by dividing the data into several sets for training and testing. It avoids overfitting, and it allows a model to generalize well to new data, gives a realistic estimate of its performance, and makes efficient use of available data. Cross-validation offers an opportunity to try out several models on the subsets to obtain the best among them, perform fine-tuning of hyperparameters, and get the measure of variation in performance over different splits of data. This ensures the robustness and reliability of the model for real-world applications.

## Step 4 : Model Selection

As part of the next step in our analysis, we use models like Linear Regression, Lasso and Ridge. All the models are further compared.

## Modeling :

### Linear Regression Model:

Cross validation was applied while training the dataset on this model.

---

```
Cross-Validation RMSE Scores: [3.83468671 4.44997437 3.98280663 4.24789734 4.07006985]  
Mean RMSE: 4.117086979594635  
Standard Deviation RMSE: 0.21350274121335028
```

```
Linear Regression Performance on Test Set:  
RMSE: 3.9042207160475693  
R2 Score: 0.8240562050040523
```

The linear regression model showed very strong performance, with an average RMSE of 4.12 and a very low standard deviation of 0.21 across cross-validation folds, indicating very consistent predictive accuracy. On the test set, the model achieved an RMSE of 3.90 and an  $R^2$  score of 0.824. With an R-squared score of 0.824, this means the model explains about 82.4% of the variance in the target variable. This shows that the linear regression model fits well and can give reliable predictions.

### Coefficients:

Year	0.022660
Status	-1.923063
Adult Mortality	-20.042525
infant deaths	0.093185
Alcohol	0.103052
percentage expenditure	6.166835
Hepatitis B	-0.016243
Measles	-0.000022
BMI	5.971843
under-five deaths	-0.069450
Polio	0.028077
Total expenditure	0.018456
Diphtheria	0.041255
HIV/AIDS	-46.304235
GDP	5.672466
Population	-0.977671
thinness 1-19 years	-0.115544
thinness 5-9 years	0.007602
Income composition of resources	6.638998
Schooling	27.220699
dtype: float64	

### Significant Predictors:

- **Schooling:** Higher years of schooling significantly increase life expectancy.
- **GDP:** Higher GDP is strongly associated with increased life expectancy.
- **Income Composition of Resources:** Better income resource composition positively influences life expectancy.
- **HIV/AIDS:** Higher HIV/AIDS death rates drastically reduce life expectancy.
- **Adult Mortality:** Increased adult mortality rates significantly lower life expectancy.
- **Under-Five Deaths:** Higher under-five mortality negatively impacts life expectancy.

## Lasso Regression :

Lasso Regression Performance:

Optimal Alpha: 0.01

RMSE (Test Set): 3.9203170646535255

R<sup>2</sup> Score (Test Set): 0.8226024497637563

Lasso Coefficients:

Schooling	25.088369
Income composition of resources	6.643154
GDP	5.869918
BMI	5.606661
percentage expenditure	3.830008
Alcohol	0.117598
infant deaths	0.097121
Diphtheria	0.042805
Year	0.033181
Polio	0.029690
Total expenditure	0.014552
Population	-0.000000
Measles	-0.000022
thinness 5-9 years	-0.004647
Hepatitis B	-0.016756
under-five deaths	-0.072314
thinness 1-19 years	-0.126445
Status	-2.090170
Adult Mortality	-20.161303
HIV/AIDS	-44.161285

dtype: float64

The Lasso regression model achieves an R<sup>2</sup> score of **0.8226**, explaining 82.26% of the variance in life expectancy. With an RMSE of **3.92**, the model demonstrates strong predictive performance and effectively identifies key predictors by penalizing less important features, enhancing model simplicity and interpretability. These results indicate a reliable fit and highlight actionable factors influencing life expectancy.

## Significant Predictors:

- **Schooling:** The most significant positive predictor, where higher years of schooling greatly improve life expectancy.
- **Income Composition of Resources:** Strongly associated with increased life expectancy, reflecting the importance of access to resources like education and income.

- **GDP:** Higher GDP significantly contributes to longer life expectancy, indicating the role of economic prosperity.
- **BMI:** A positive association with life expectancy, reflecting the importance of nutrition and health.
- **Percentage Expenditure:** Increased health expenditure is linked to improved life expectancy.
- **HIV/AIDS:** The most significant negative predictor, with higher HIV/AIDS mortality drastically reducing life expectancy.
- **Adult Mortality:** A strong negative influence, where higher adult mortality rates lower life expectancy.
- **Under-Five Deaths:** Negatively impacts life expectancy, highlighting the importance of child health.
- **Status:** Developing countries (negative impact) have shorter life expectancy compared to developed countries.

## Ridge Regression:

Ridge Regression Performance:

Optimal Alpha: 0.1

RMSE (Test Set): 3.9142660292171754

R<sup>2</sup> Score (Test Set): 0.8231496557244495

Ridge Coefficients:

Schooling	26.761544
Income composition of resources	6.788495
percentage expenditure	6.082255
BMI	6.013748
GDP	5.727232
Alcohol	0.103708
infant deaths	0.093730
Diphtheria	0.041398
Polio	0.028234
Year	0.023960
Total expenditure	0.017555
thinness 5-9 years	0.006934
Measles	-0.000022
Hepatitis B	-0.016239
under-five deaths	-0.069849
thinness 1-19 years	-0.116717
Population	-0.853351
Status	-1.939776
Adult Mortality	-20.120228
HIV/AIDS	-45.788400

dtype: float64

## Significant Predictors:

- **Schooling:** The most significant positive predictor, where more years of education strongly improve life expectancy.
- **Income Composition of Resources:** A crucial positive factor, showing that better resource distribution significantly increases life expectancy.
- **Percentage Expenditure:** Higher government health expenditure is positively associated with improved life expectancy.
- **BMI:** Indicates that better nutrition and health positively influence life expectancy.
- **GDP:** Economic prosperity contributes significantly to longer life expectancy.
- **HIV/AIDS:** The strongest negative predictor, with higher mortality from HIV/AIDS drastically reducing life expectancy.
- **Adult Mortality:** Higher adult mortality rates negatively affect life expectancy.



- **Under-Five Deaths:** Higher child mortality reduces life expectancy, emphasizing the importance of early childhood health.
- **Status:** Developing countries (negative impact) have shorter life expectancy compared to developed countries.

## Summary:

The Ridge regression model, with an optimal alpha of 0.1, achieves an  $R^2$  score of **0.8231**, explaining 82.31% of the variance in life expectancy. With an RMSE of **3.91**, the model demonstrates excellent predictive performance while effectively handling multicollinearity through L2 regularization, ensuring all features contribute meaningfully without being eliminated. These results indicate a well-balanced model that provides reliable predictions and captures critical factors influencing life expectancy.

## Comparison of Models:

1. **Linear Regression:**
  - $R^2$ : 0.8241
  - RMSE: 3.90
  - Strength: Simple and interpretable but sensitive to multicollinearity.
  - Limitation: Does not handle multicollinearity or overfitting effectively.
2. **Lasso Regression:**
  - $R^2$ : 0.8226
  - RMSE: 3.92
  - Strength: Performs feature selection by shrinking coefficients of less important features to zero.
  - Limitation: May exclude some features completely, potentially losing valuable information.
3. **Ridge Regression:**
  - $R^2$ : 0.8231
  - RMSE: 3.91
  - Strength: Handles multicollinearity effectively by penalizing large coefficients while retaining all features.
  - Limitation: Does not reduce coefficients to zero, so all features are included, even minor ones.

## Conclusion:

Ridge regression is the best model for this analysis due to its balance between predictive performance and feature retention. With an  $R^2$  of 0.8231 and RMSE of 3.91, it effectively captures the relationship between all predictors and life expectancy, considers every feature which is not so in the other models considered here like Linear and Lasso while mitigating multicollinearity, making it robust and reliable for real-world applications. For a dataset

considered here which is a real world data and healthcare domain specific where all features contribute in a way or another, ridge seems the best model for analysis.

## Results

The results section presents the outcomes of data analysis, feature importance, and model performance for predicting life expectancy.

### Exploratory Insights:

- Education emerged as a key factor, with higher years of schooling significantly associated with longer life expectancy.
- Economic factors like GDP and healthcare expenditure positively influenced life expectancy.
- Poor health outcomes, such as high adult mortality, under-five mortality, and HIV/AIDS deaths, were strongly linked to shorter life expectancy.

### Feature Importance:

- **Top Predictors:** Schooling, income composition of resources, GDP, and healthcare expenditure.
- **Negative Predictors:** Adult mortality, HIV/AIDS, and under-five deaths had the strongest adverse effects on life expectancy.
- **Weaker Predictors:** Features like alcohol consumption and total expenditure showed smaller, but still relevant, associations.

## **Challenges faced:**

### Feature Relevance:

Some features (e.g., Population, Measles) had near-zero impact and were redundant, complicating initial analysis.

### Data Scaling:

Normalization was essential to balance feature magnitudes (e.g., GDP vs. BMI) to avoid biased predictions.

### Data Quality:

Missing values or limited feature diversity (e.g., no environmental data) restricted the model's scope.

## **Improvement suggestions:**

### Economic Indicators:

Add unemployment rate or income inequality indices for deeper economic analysis.

### Healthcare Infrastructure:

Include hospital density, access to healthcare, and health expenditure per capita.

### Environment and Lifestyle:

Add air quality indices, smoking rates, and dietary habits for better insights into health and environmental factors.

### Regional/Geographical Factors:

Regional differences (e.g., rural vs. urban) may improve granularity in predictions.

