LIFE EXPECTANCY

SINDHURA SURAVARJHULA
MRUNALI SAROJ NILGIRWAR

UNDER THE GUIDANCE OF
DR. MEI YANG

ABSTRACT

This analysis explores a dataset focused on factors influencing life expectancy across countries. The dataset includes variables such as **Schooling, GDP, Income Composition of Resources, Adult Mortality, and healthcare-related** indicators. The primary objective was to **predict life expectancy using regression models** and identify the most impactful predictors.

OBJECTIVE

The primary goal of this project is to analyze and predict the factors that significantly influence life expectancy using a comprehensive dataset.

AIM

- **Predict Life Expectancy** : Build regression models to predict life expectancy based on socio-economic and health-related indicators.
- **Understand Feature Relationships** : Explore relationships among variables through correlation analysis, scatter plots etc.
- **Identify Key Predictors** : Identify the most critical factors affecting life expectancy using feature selection methods and models.
- **Compare Model Performance** : Evaluate the performance of multiple regression models (Linear, Ridge, Lasso) to determine the most effective approach for accurate prediction.

THE DATA

Target Variable – “ Life Expectancy ”

– the average number of years a person is expected to live

The dataset contains health-related indicators and socio-economic factors from various countries, aimed at analyzing their impact on Life Expectancy.

1. **Total Rows: 2938**
2. **Total Columns: 22**

Got the Dataset from kaggle.com

THE DATA

Numerical Variables:

- Life Expectancy (Target Variable)
- Adult Mortality
- Infant Deaths
- Alcohol
- Percentage Expenditure
- Hepatitis B
- Measles
- BMI
- Under-five Deaths
- Polio
- Total Expenditure
- Diphtheria
- HIV/AIDS
- GDP
- Population
- Thinness 10-19 Years
- Thinness 5-9 Years
- Income Composition of Resources
- Schooling

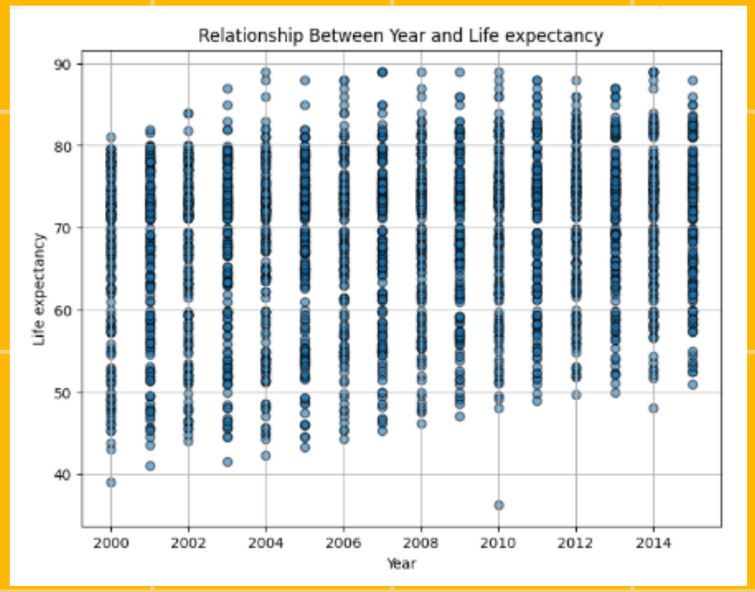
Categorical Variables:

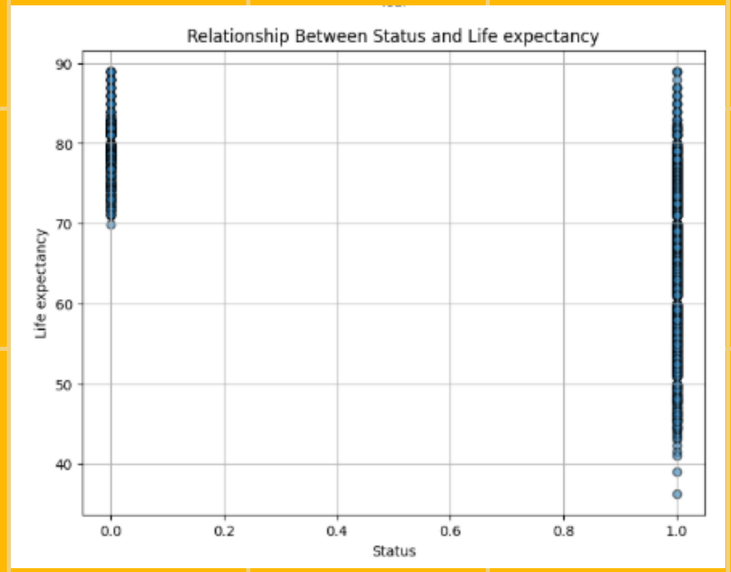
1. Country
2. Year
3. Status

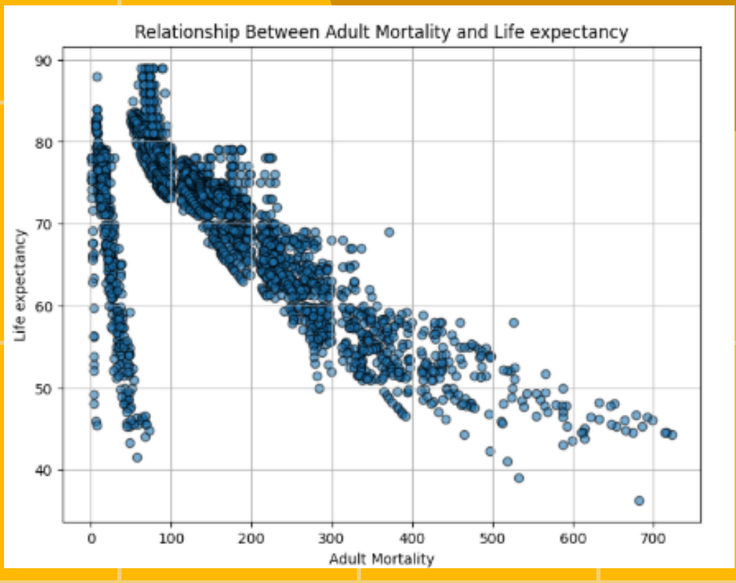
NULL VALUES

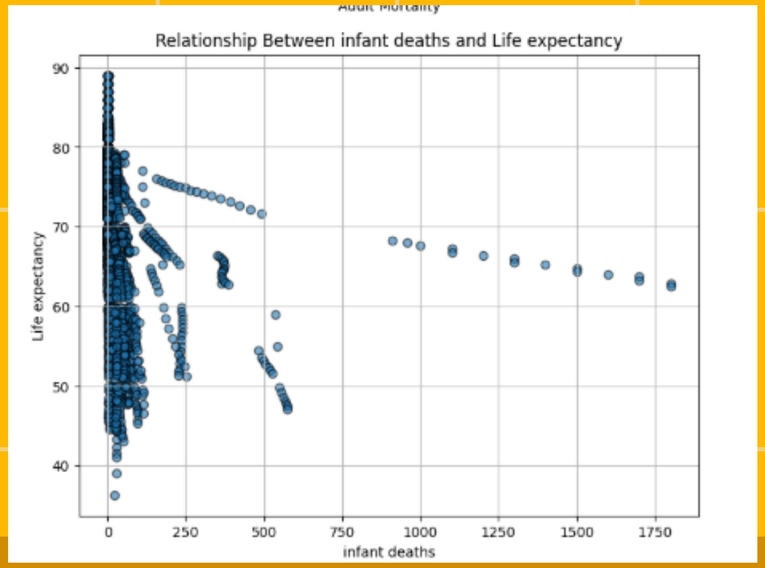
Country	0	Total Expenditure	226
Year	0	Diphtheria	19
Status	0	HIV/AIDS	0
Life Expectancy	10	GDP	448
Adult Mortality	10	Populations	652
Infant deaths	0	thinness 10-19 yrs	34
Alcohol	194	thinness 10-19 yrs	34
Percentage expenditure	0	Income composition of resources	167
Hepatitis B	553	Schooling	163
Measles	0		
BMI	34		
under-five deaths	0		
Polio	19		

PLOTS

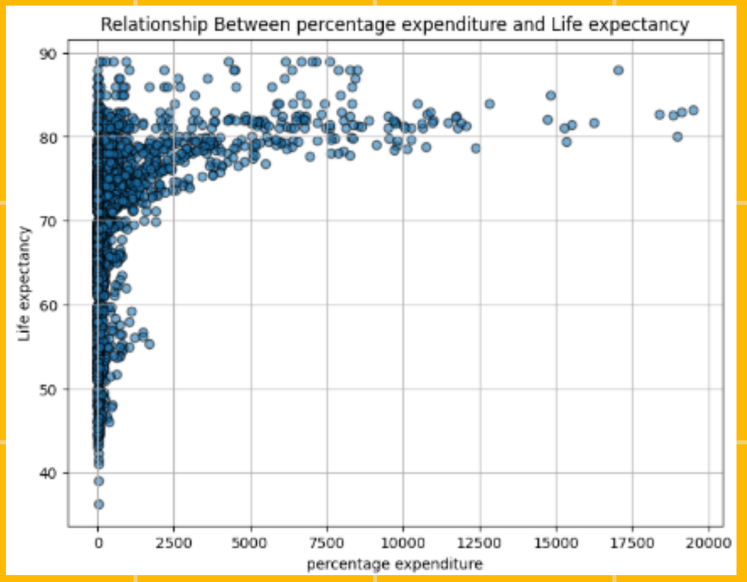

Relationship Between Year and Life expectancy
Life expectancy
Year
2000
2002
2004
2006
2008
2010
2012
2014
40
50
60
70
80
90


Relationship Between Status and Life expectancy
Life expectancy
Status
0.0
0.2
0.4
0.6
0.8
1.0
40
50
60
70
80
90

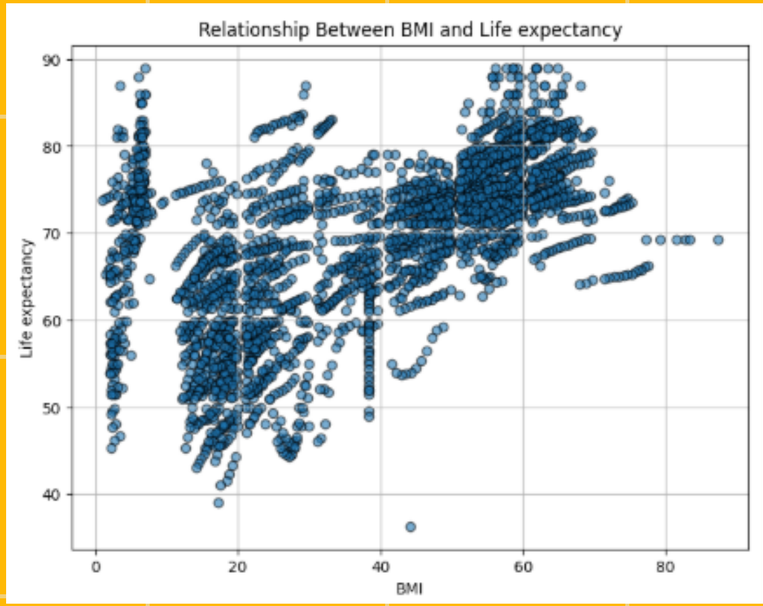

Relationship Between Adult Mortality and Life expectancy
Life expectancy
Adult Mortality
0
100
200
300
400
500
600
700
40
50
60
70
80
90

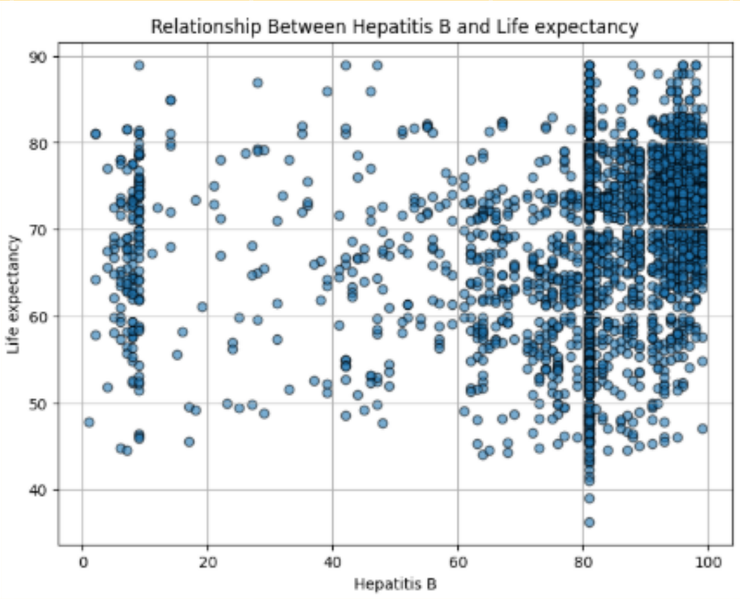

Relationship Between infant deaths and Life expectancy
Life expectancy
infant deaths
0
250
500
750
1000
1250
1500
1750
40
50
60
70
80
90

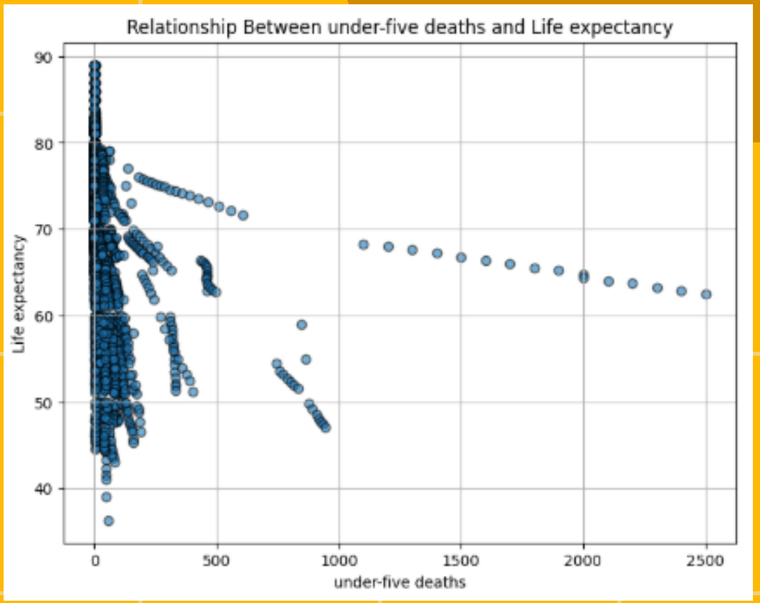

Relationship Between Alcohol and Life expectancy
Life expectancy
Alcohol
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
40
50
60
70
80
90

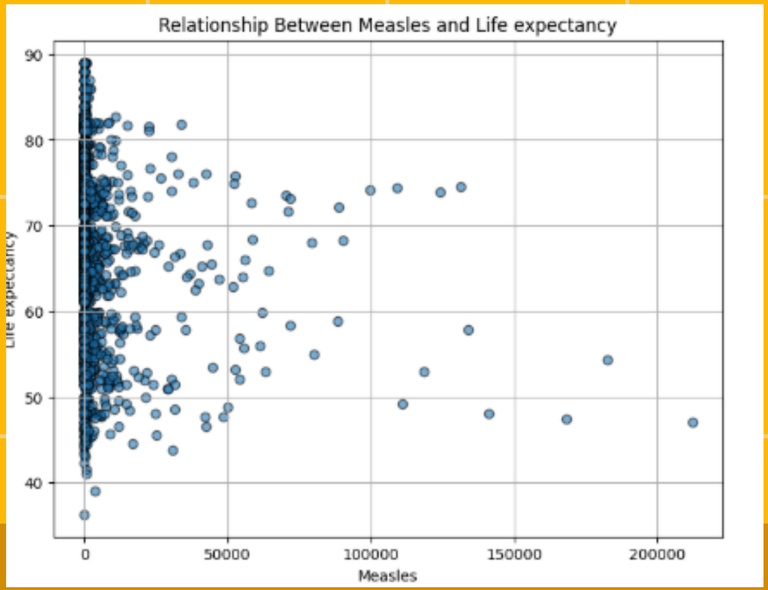

Relationship Between percentage expenditure and Life expectancy
Life expectancy
percentage expenditure
0
2500
5000
7500
10000
12500
15000
17500
20000
40
50
60
70
80
90

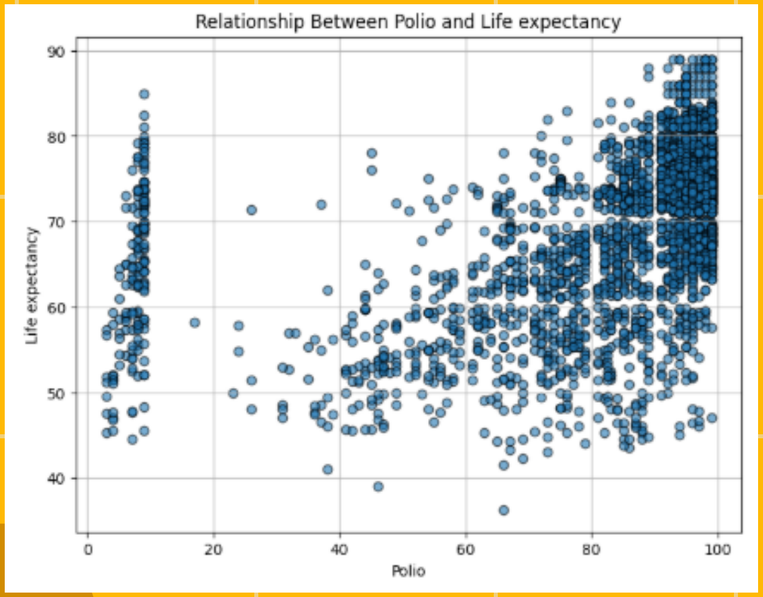
PLOTS

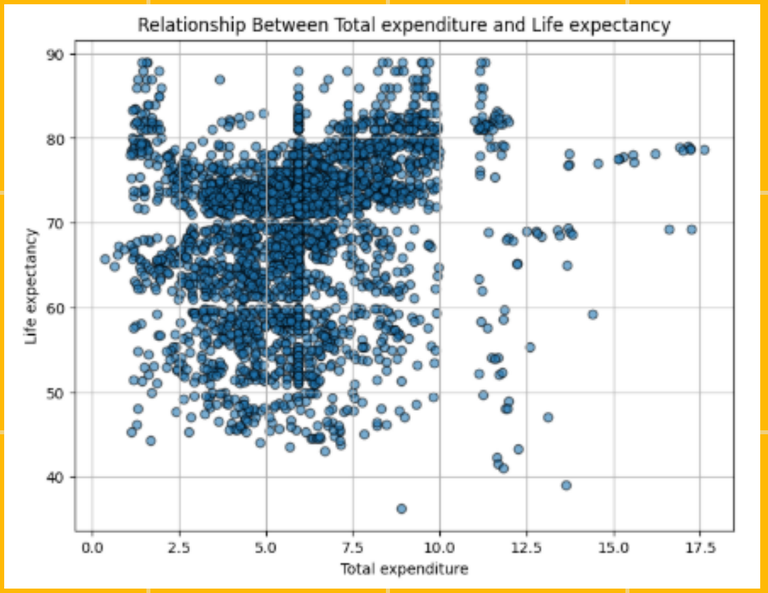

Relationship Between BMI and Life expectancy
Life expectancy
BMI


Relationship Between Hepatitis B and Life expectancy
Life expectancy
Hepatitis B

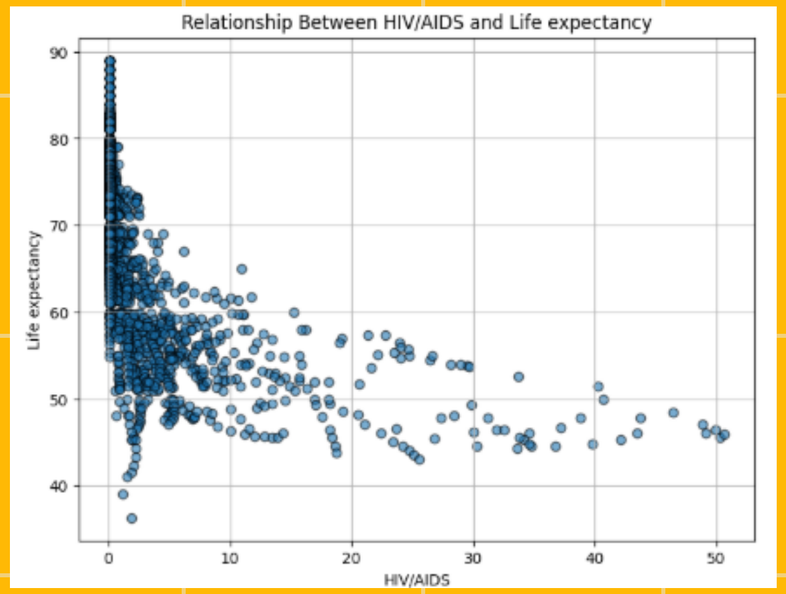

Relationship Between under-five deaths and Life expectancy
Life expectancy
under-five deaths

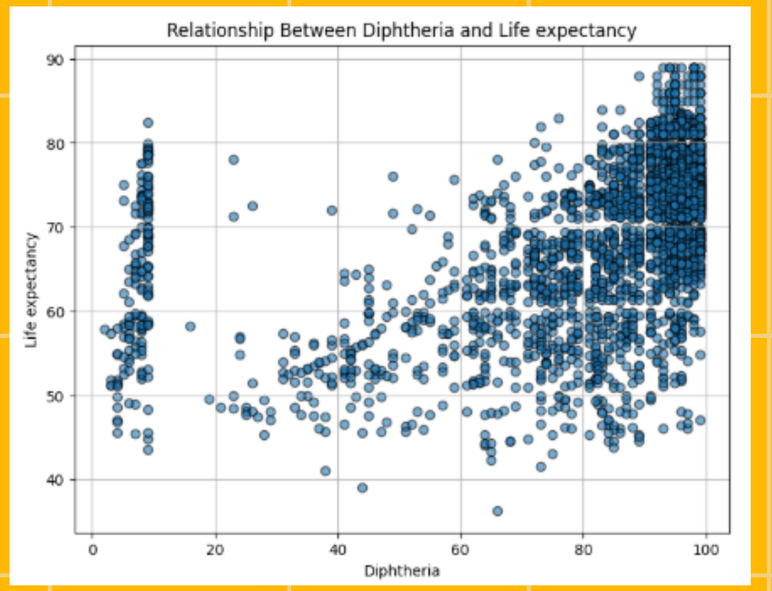

Relationship Between Measles and Life expectancy
Life expectancy
Measles

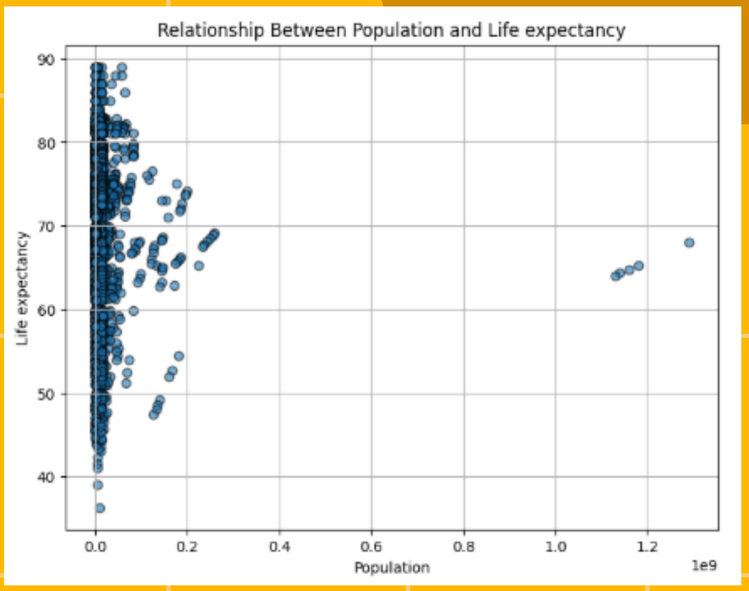

Relationship Between Polio and Life expectancy
Life expectancy
Polio

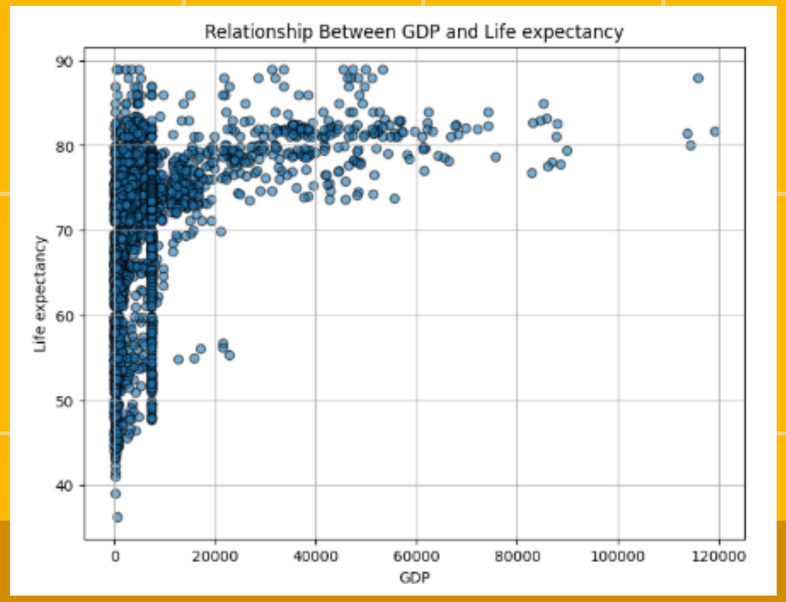

Relationship Between Total expenditure and Life expectancy
Life expectancy
Total expenditure

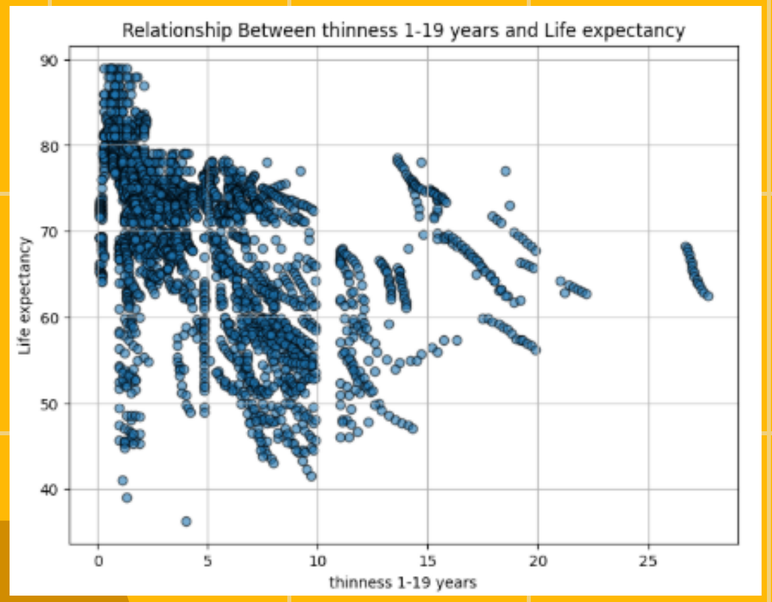
PLOTS


Relationship Between HIV/AIDS and Life expectancy
Life expectancy
HIV/AIDS


Relationship Between Diphtheria and Life expectancy
Life expectancy
Diphtheria

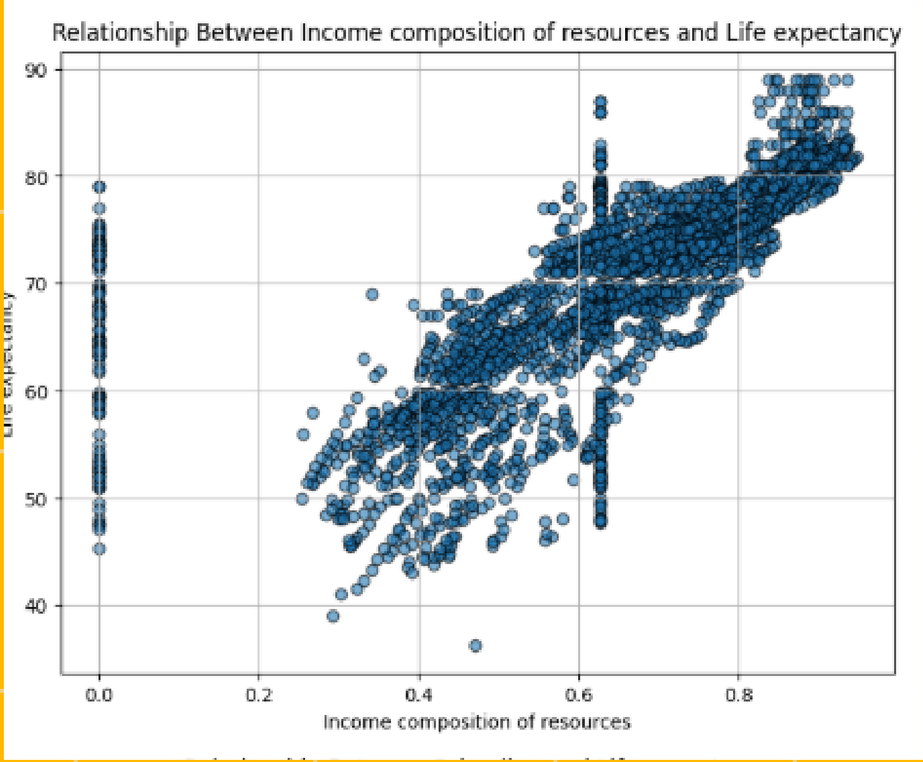

Relationship Between Population and Life expectancy
Life expectancy
Population
1e9


Relationship Between GDP and Life expectancy
Life expectancy
GDP


Relationship Between thinness 1-19 years and Life expectancy
Life expectancy
thinness 1-19 years


Relationship Between thinness 5-9 years and Life expectancy
Life expectancy
thinness 5-9 years

PLOTS

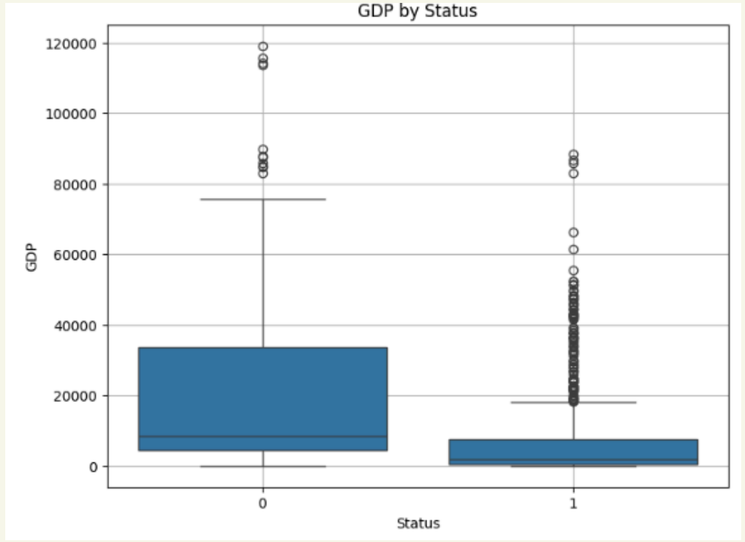

Relationship Between Income composition of resources and Life expectancy
Life expectancy
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8
Income composition of resources

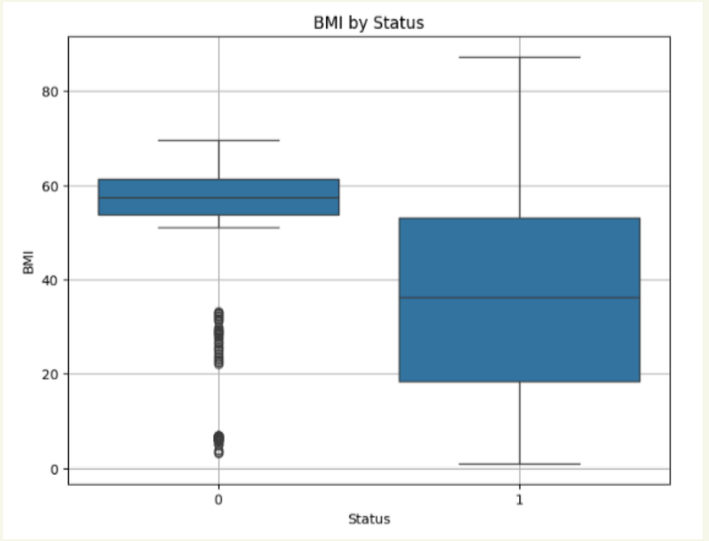

Relationship Between Schooling and Life expectancy
Life expectancy
90
80
70
60
50
40
0
5
10
15
20
Schooling

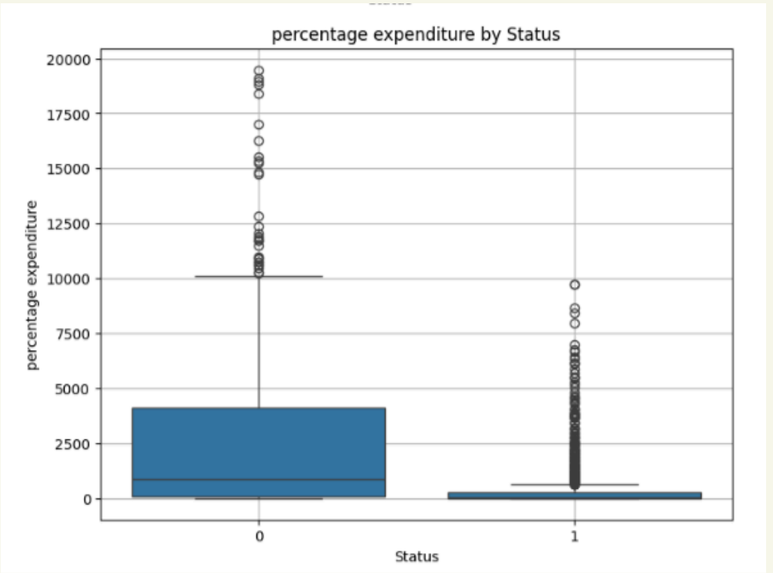
CONCLUSIONS FROM PLOTS

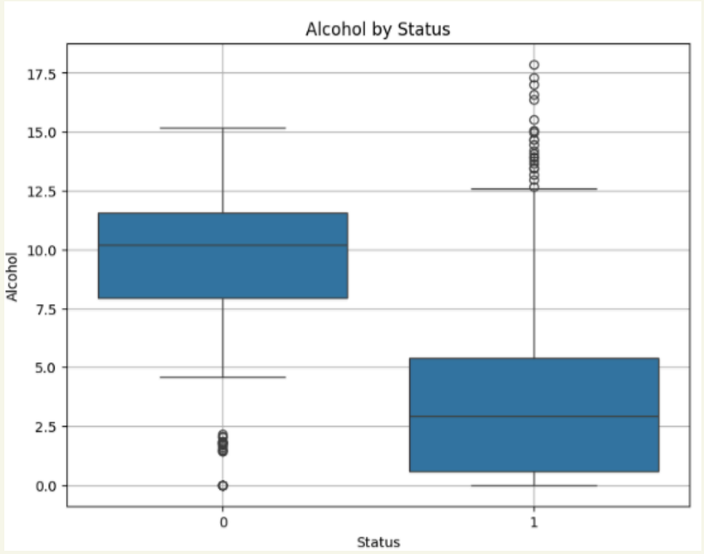
- **Strongest Predictors:**
  - **Adult Mortality**, **Under-Five Deaths**, **Income Composition of Resources**, **Schooling**, and **GDP** are the most significant predictors of life expectancy.
  - These variables have clear and consistent relationships with life expectancy, making them critical for modeling and interpretation.
- **Moderate Predictors:**
  - **Percentage Expenditure**, **BMI**, and **immunization rates** (Diphtheria, Polio) contribute moderately to life expectancy but are less consistent compared to strong predictors.
- **Weak Predictors:**
  - **Alcohol Consumption, Measles**, and **Population size** show no strong trends and likely contribute minimally to life expectancy variation.

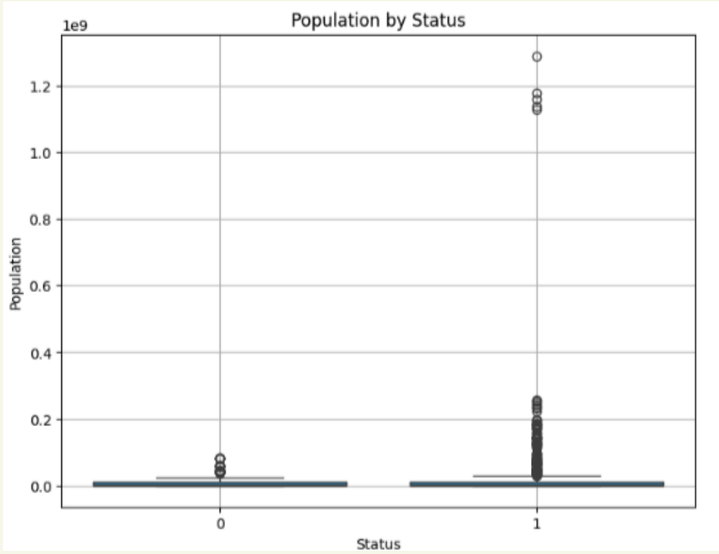
PLOTS


GDP by Status
GDP
Status
0
20000
40000
60000
80000
100000
120000
0
1


BMI by Status
BMI
Status
0
20
40
60
80
0
1


percentage expenditure by Status
percentage expenditure
Status
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1


Alcohol by Status
Alcohol
Status
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
1


Population by Status
1e9
Population
Status
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1

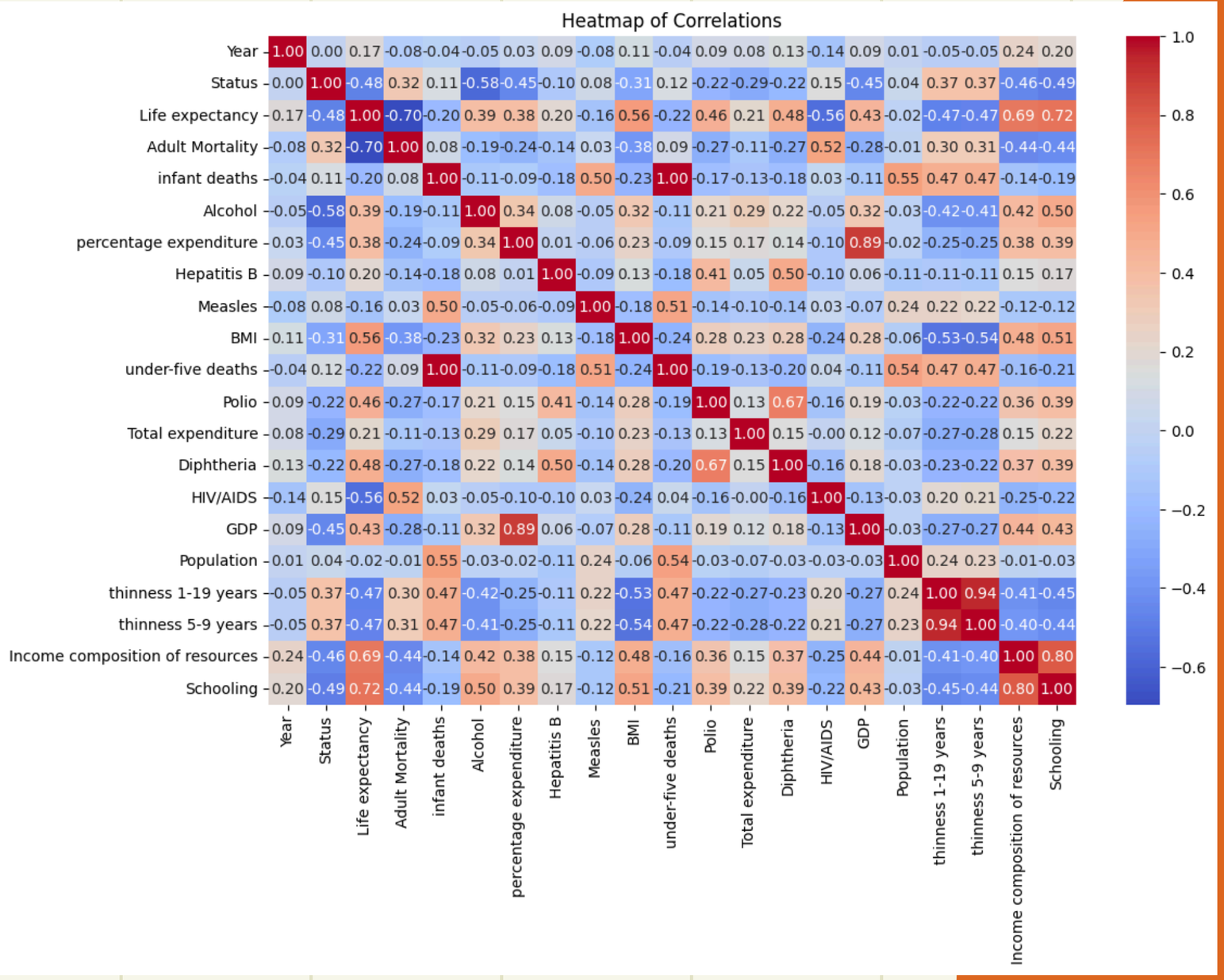
CONCLUSIONS FROM BOX PLOTS

- Healthcare Indicators (BMI, Expenditure) :
  - Developed countries have better health-related distributions (higher BMI, more efficient healthcare spending).

- Economic Factors (GDP, Population) :
  - GDP and population size alone do not differentiate development status clear

- Lifestyle (Alcohol) :
  - Higher alcohol consumption in developing countries suggests potential cultural or economic differences in behavior.

CORRELATION HEATMAP



CONCLUSIONS FROM HEATMAP

- **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.

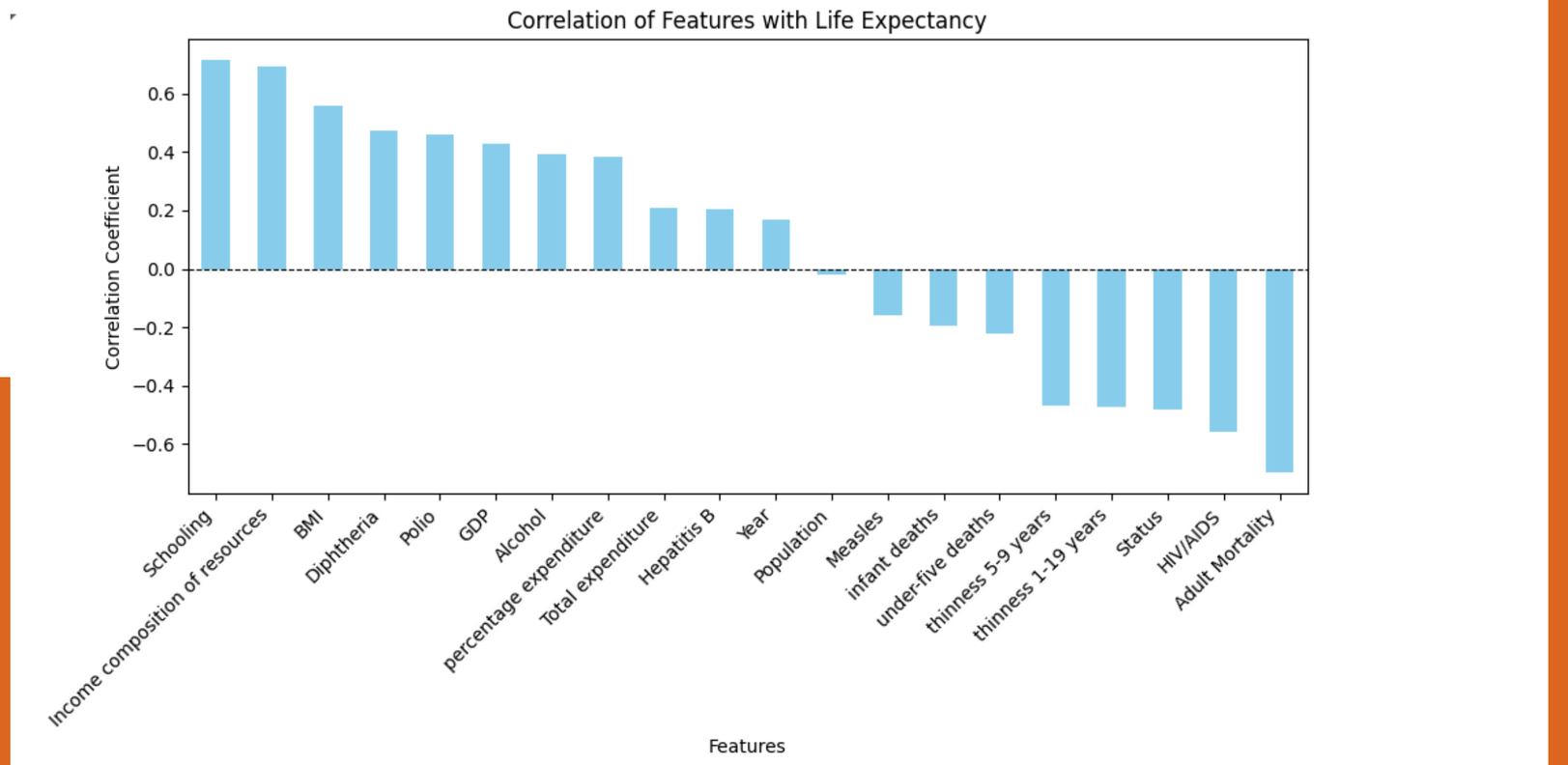
Heatmap Observations

- 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.

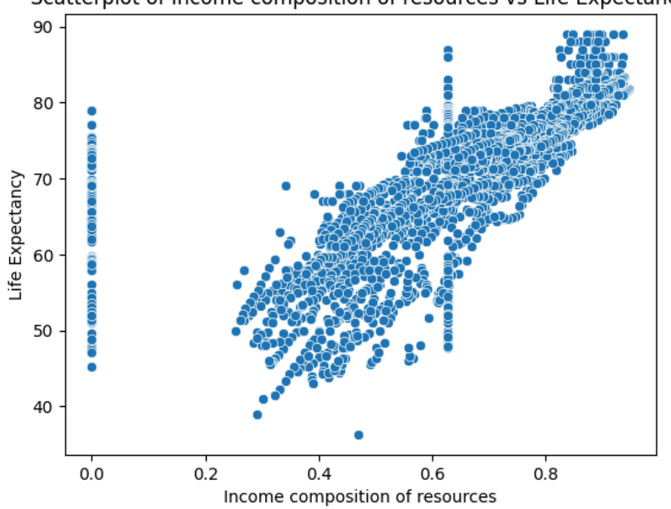
LINEAR ASSOCIATIONS

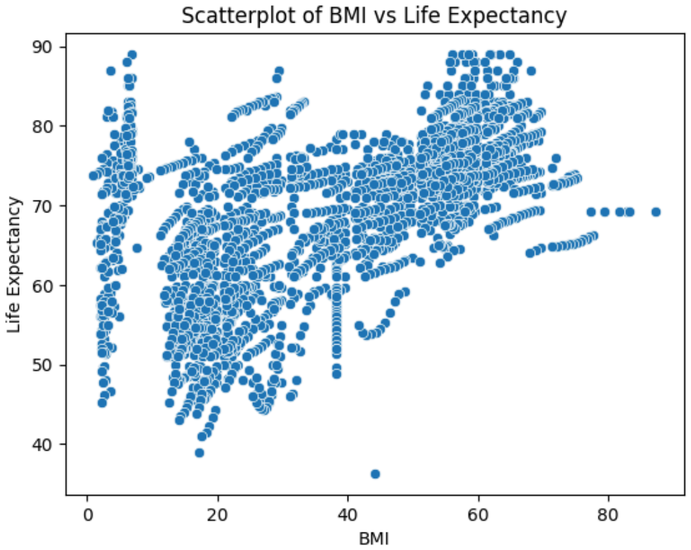
```
Linear Associations (Correlations) with Life Expectancy:
Schooling                          0.715066
Income composition of resources    0.692483
BMI                                0.559255
Diphtheria                         0.475418
Polio                              0.461574
GDP                                0.430493
Alcohol                            0.391598
percentage expenditure             0.381791
Total expenditure                  0.207981
Hepatitis B                        0.203771
Year                               0.169623
Population                        -0.019640
Measles                           -0.157574
infant deaths                     -0.196535
under-five deaths                 -0.222503
thinness 5-9 years                -0.466629
thinness 1-19 years               -0.472162
Status                            -0.481962
HIV/AIDS                          -0.556457
Adult Mortality                   -0.696359
dtype: float64
```

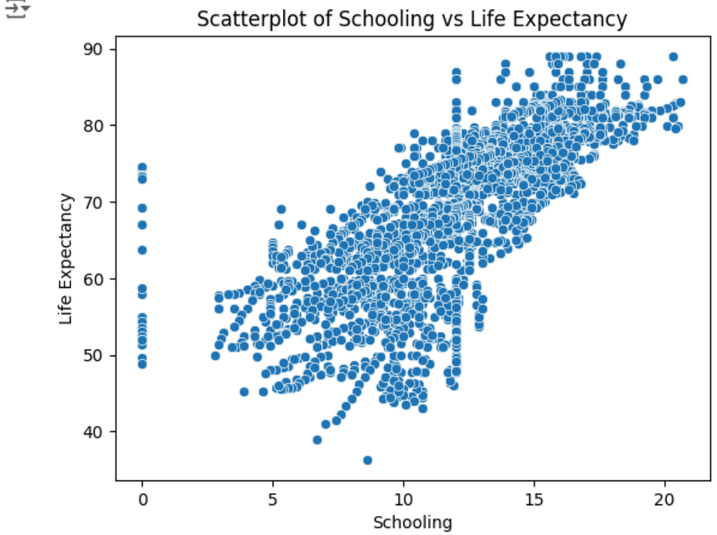
dtype: float64



Schooling


Scatterplot of Income composition of resources vs Life Expectancy
Life Expectancy
Income composition of resources
90
80
70
60
50
40
0.0
0.2
0.4
0.6
0.8


Scatterplot of BMI vs Life Expectancy
Life Expectancy
BMI
90
80
70
60
50
40
0
20
40
60
80


Scatterplot of Schooling vs Life Expectancy
Life Expectancy
Schooling
90
80
70
60
50
40
0
5
10
15
20

CONCLUSIONS OBSERVED

- **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.

CONCLUSIONS OBSERVED

- **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.

CONCLUSIONS OBSERVED

- **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.

CONCLUSIONS OBSERVED

- **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.

LINEAR REGRESSION

```
  warnings.warn(
Cross-Validation RMSE Scores: [3.77964668 4.52993127 4.03903617 4.36487012 4.05392044]
Mean RMSE: 4.153480935548676
Standard Deviation RMSE: 0.26425828047348393

Linear Regression Performance on Test Set:
RMSE: 3.913688107973489
R² Score: 0.8232018739564674
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_regression.py:492: FutureWarning: 'squared' is deprecated in version 1.4 a
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_regression.py:492: FutureWarning: 'squared' is deprecated in version 1.4 a
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_regression.py:492: FutureWarning: 'squared' is deprecated in version 1.4 a
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_regression.py:492: FutureWarning: 'squared' is deprecated in version 1.4 a
  warnings.warn(
```

Model Performance:

Cross-Validation Mean RMSE: 4.15

Test Set RMSE: 3.91

Test Set R2 score: 82.32%

"The model predicts life expectancy with an average error of 3.91 years, explaining 82.32% of the variance, showcasing high accuracy and robustness."

Consistency Across Folds:

- Cross-Validation Standard Deviation (RMSE): 0.26

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	

LASSO REGRESSION

```
  model = cd_fast.enet_coordinate_descent_gram(
Lasso Regression Performance:
Optimal Alpha: 0.01
RMSE (Test Set): 3.91037686890478
R² Score (Test Set): 0.8235009132225997

Lasso Coefficients:
Schooling                          19.053910
Income composition of resources     6.511180
GDP                                 4.913044
BMI                                 4.531283
percentage expenditure              2.047078
infant deaths                       0.097350
Alcohol                             0.093340
Diphtheria                          0.042001
Polio                               0.029009
Total expenditure                   0.022069
Year                                0.020125
Population                         -0.000000
thinness 5-9 years                 -0.000000
Measles                            -0.000022
Hepatitis B                        -0.017179
under-five deaths                  -0.072486
thinness 1-19 years                -0.113624
Status                             -1.936866
Adult Mortality                   -17.768298
HIV/AIDS                          -33.232423
dtype: float64
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_
  model = cd_fast.enet_coordinate_descent(
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_regre
```

- **Optimal Alpha: 0.01 (selected through cross-validation)**
- **Test Set Performance:**
  - **RMSE: 3.91**
  - **R^2 Score: 82.35%**

- Lasso regression effectively performed feature selection, identifying key predictors such as Schooling, Income Composition of Resources, and GDP as the most impactful positive contributors, while reducing less significant features to zero."
- Lasso regression predicts life expectancy with an average error of 3.91 years and explains 82.35% of the variance in the test data.

RIDGE REGRESSION

```
Ridge Regression Performance:
Optimal Alpha: 0.1
RMSE (Test Set): 3.9042225312923136
R² Score (Test Set): 0.8240560413959231

Ridge Coefficients:
Schooling                          20.029114
Income composition of resources     6.489438
GDP                                 4.808830
BMI                                 4.777123
percentage expenditure              3.701948
infant deaths                       0.095211
Alcohol                             0.082201
Diphtheria                          0.040968
Polio                               0.027912
Total expenditure                   0.024141
Year                                0.013131
thinness 5-9 years                  0.012651
Measles                            -0.000022
Hepatitis B                        -0.016831
under-five deaths                  -0.070911
thinness 1-19 years                -0.108781
Population                         -0.953584
Status                             -1.830621
Adult Mortality                   -17.695652
HIV/AIDS                          -34.498300
dtype: float64
/usr/local/lib/python3.10/dist-packages/sklearn/
  warnings.warn(
```

Model Performance:

Cross-Validation Mean RMSE: 4.13

Test Set RMSE: 3.90

Test Set R2: 82.41%

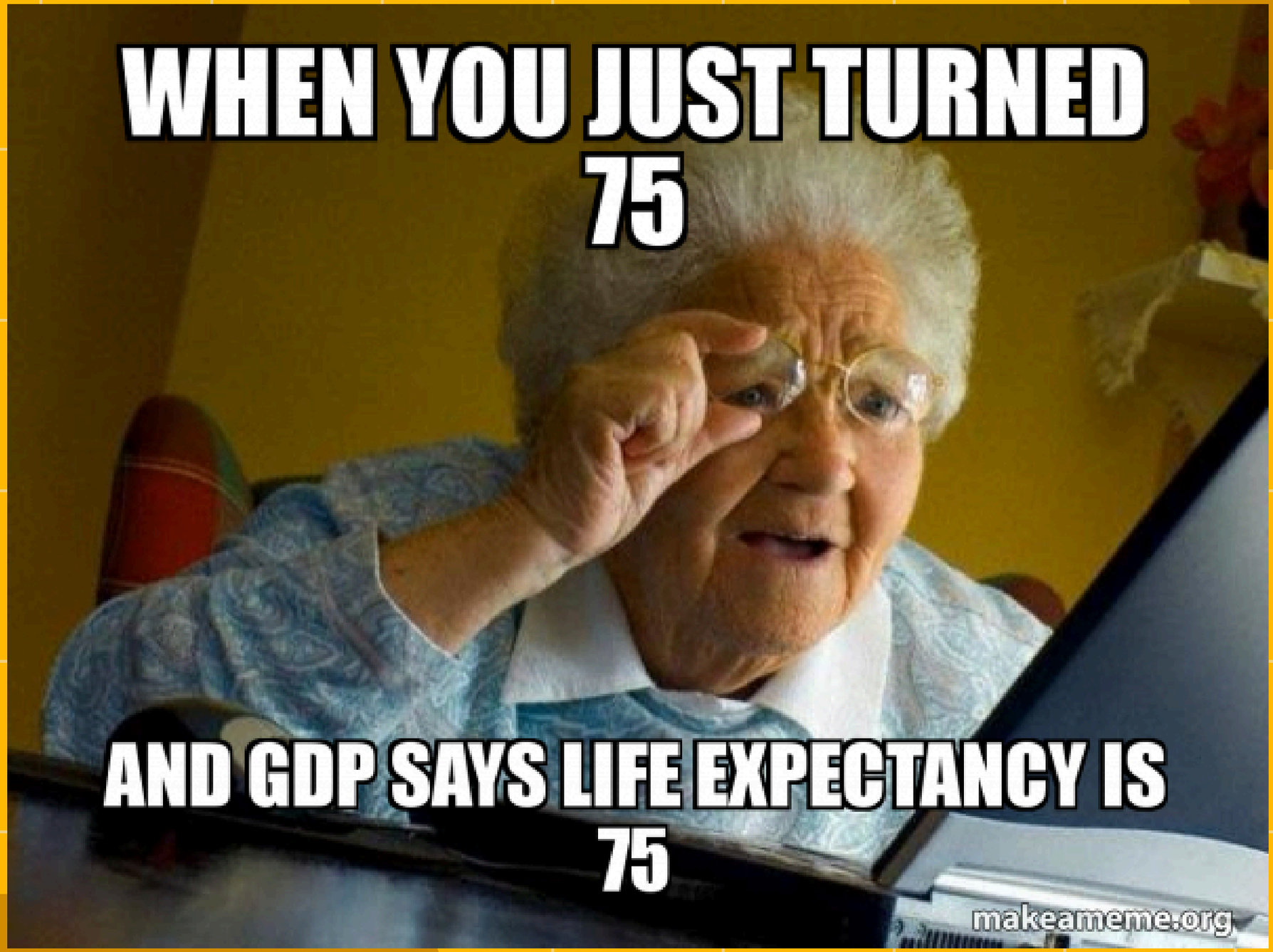
- Unlike Lasso, Ridge regression does not shrink any coefficients to zero, making it suitable for scenarios where all features are considered important but need to be regularized.
- Ridge regression achieves a strong prediction performance with an average error of 3.90 years and explains 82.41% of the variance in life expectancy on the test set.

COMPARISION BETWEEN MODELS

Overall Performance:

- All three models deliver similar performance, with an RMSE of ~3.90 and R2 of ~82.41%.
- This indicates that all models are equally capable of predicting life expectancy accurately.

- In real-world problems like life expectancy, even minor features (e.g., Population, Alcohol Consumption) can provide valuable insights, especially when relationships between predictors are subtle. Ridge Regression addresses multicollinearity (e.g., Schooling and Income Composition of Resources) by stabilizing predictions without discarding features. Retaining all features ensures robust predictions across diverse scenarios, including outliers and edge cases.


WHEN YOU JUST TURNED 75
AND GDP SAYS LIFE EXPECTANCY IS 75
makeameme.org

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.

IMPROVEMENTS

1. **Economic Indicators:**
   - Add unemployment rate or income inequality indices for deeper economic analysis.
2. **Healthcare Infrastructure:**
   - Include hospital density, access to healthcare, and health expenditure per capita.
3. **Environment and Lifestyle:**
   - Add air quality indices, smoking rates, and dietary habits for better insights into health and environmental factors.
4. **Regional/Geographical Factors:**
   - Regional differences (e.g., rural vs. urban) may improve granularity in predictions.

THANK YOU

QUESTIONS ?