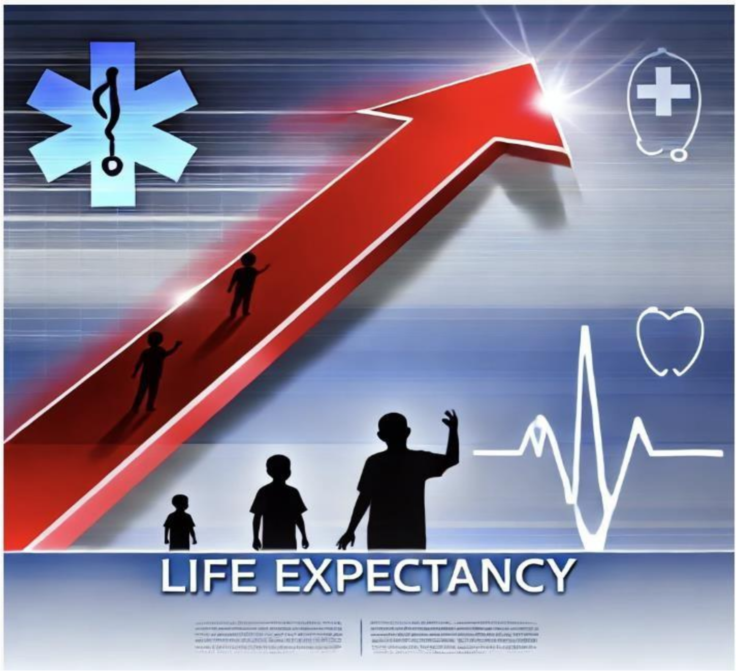
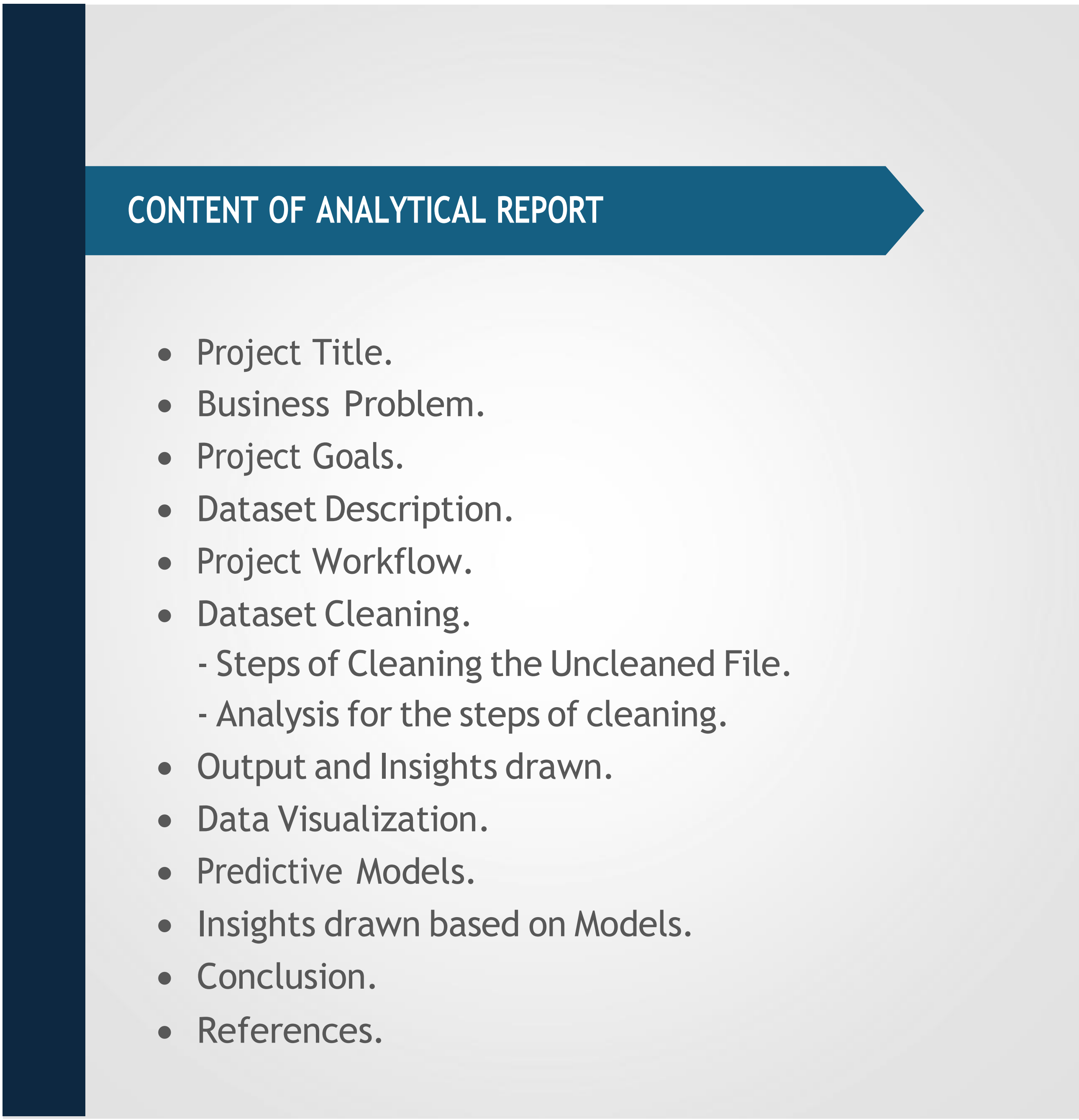

LIFE EXPECTANCY



CONTENT OF ANALYTICAL REPORT

PROJECT TITLE

LIFE EXPECTANCY ANALYSIS

Factors Inffuencing Global Health Outcomes

BUSINESS PROBLEM

Our objective: Identify key factors influencing life expectancy and provide actionable insights for policymakers and public health organizations.

PROJECT GOALS

Enhance Data Quality: Ensure accuracy by handling missing values, outliers, and inconsistencies for reliable analysis.

Enable Analysis s Insights: Prepare data for identifying key factors influencing life expectancy and provide actionable insights to policymakers.

Improve Model s Outcomes: Support robust visualizations, better model performance, and reproducible results for targeted health and economic interventions.

DATASET DESCRIPTION

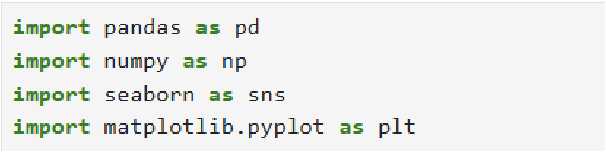
The dataset provides country-wise data over multiple years on health, economic, and demographic factors influencing life expectancy. **Key features include health indicators (e.g., mortality rates, immunization, BMI), economic metrics (GDP, healthcare expenditure), and demographics (population, schooling, income).** The target variable is life expectancy, aimed at identifying its key influencers.

PROJECT WORKFLOW

1. Import necessary libraries.
2. Load the dataset into a Pandas DataFrame.
3. Conduct a sanity check of the data:
    - Examine its structure, missing values, duplicates, and garbage values.
4. Perform exploratory data analysis (EDA) to understand distributions, relationships, and trends.
5. Handle missing values, outliers, and encode categorical variables.
6. Prepare the dataset for predictive modeling (linear regression).
7. Document insights and findings.

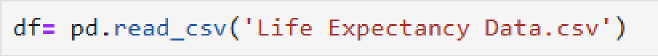
Let's look at steps, we followed in Jupyter notebook, for cleaning our dataset.

1. **Import Libraries**



```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

2. **Load Uncleaned Dataset into pandas**

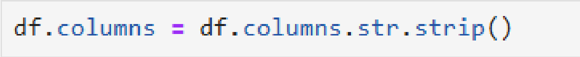


```
df= pd.read_csv('Life Expectancy Data.csv')
```

3. **Preview the First Few Rows (df. head()), and Preview the Last Few Rows (df.tail())**

4. **Strip Whitespaces from Column Names:**

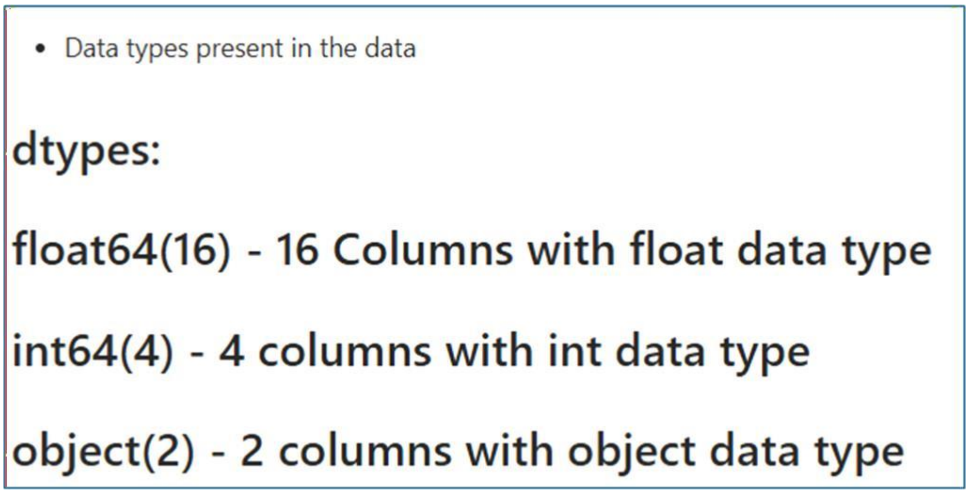
   - Ensuring column names are clean and standardized for easier access and analysis.



```
df.columns = df.columns.str.strip()
```

5. **Let's, Inspect Dataset Information**

   Provides an overview of the dataset, including column names, non-null counts, and data types.



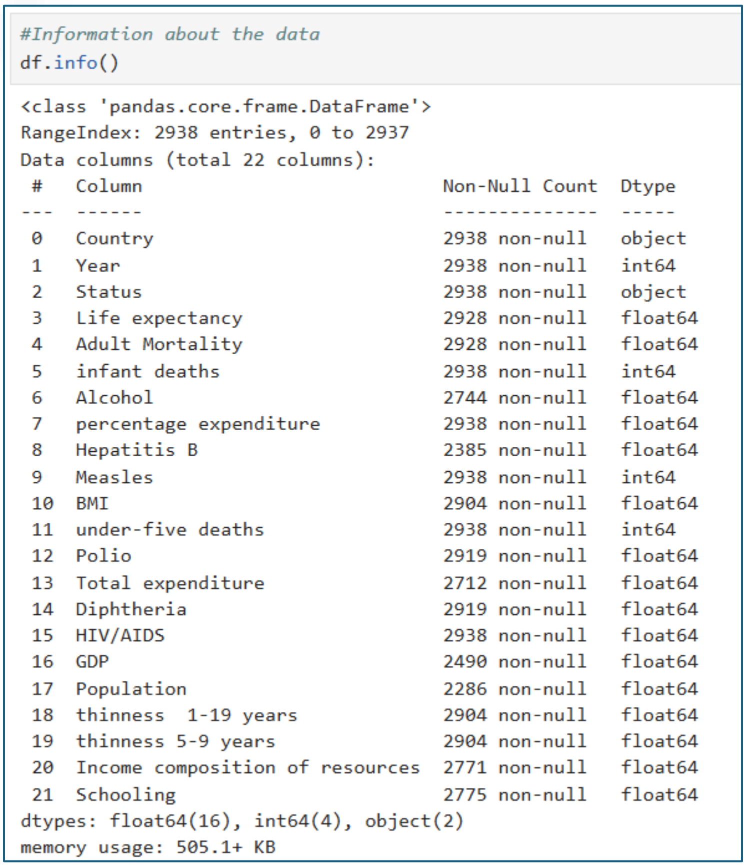
- Data types present in the data

dtypes:

float64(16) - 16 Columns with float data type

int64(4) - 4 columns with int data type

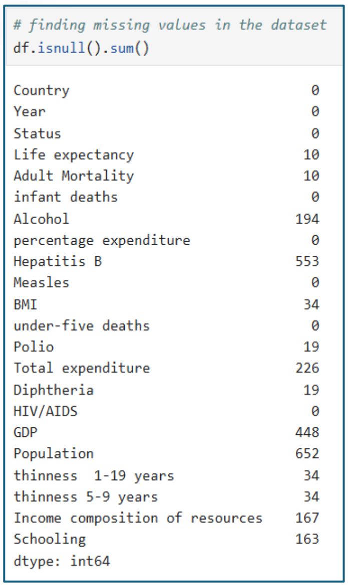
object(2) - 2 columns with object data type



```
#Information about the data
df.info()
```

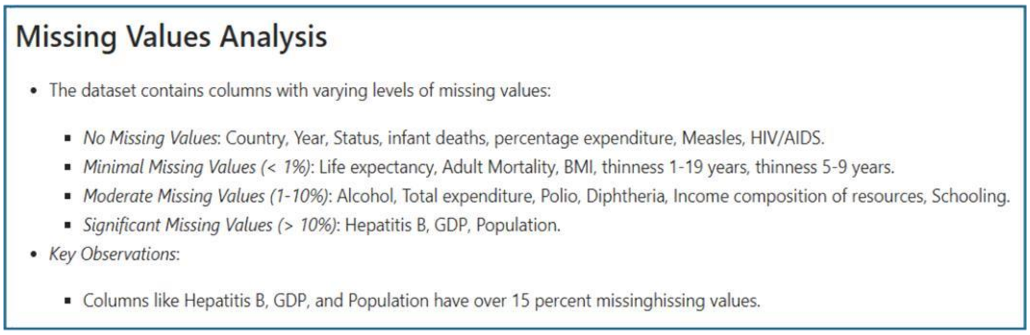
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          2938 non-null   object 
 1   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
```

6. **Check for Missing Values:** Prior to conducting any data analysis or additional Python cleaning, null values must be fixed because missing data might introduce mistakes, skew results, or cause functions to fail.



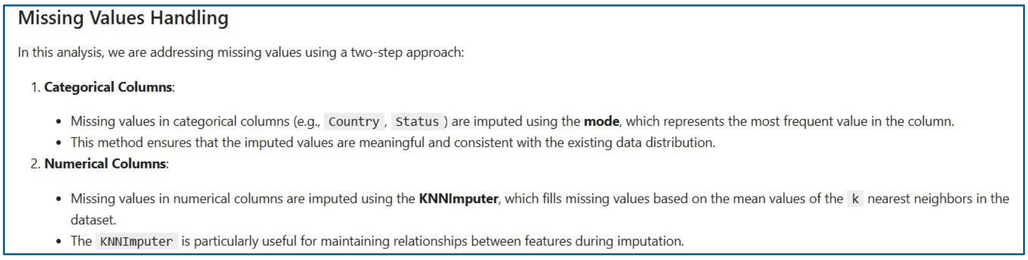
```
# finding missing values in the dataset
df.isnull().sum()
```

```
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
```



Missing Values Analysis

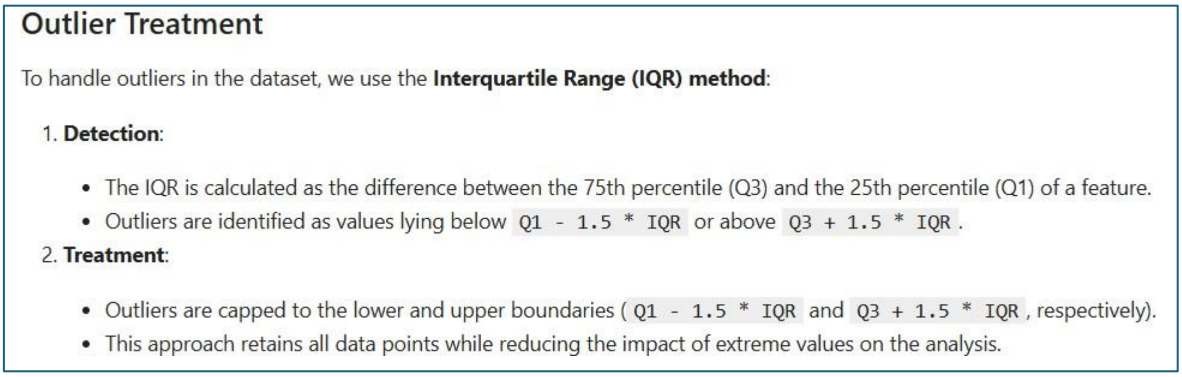
- The dataset contains columns with varying levels of missing values:
  - *No Missing Values*: Country, Year, Status, infant deaths, percentage expenditure, Measles, HIV/AIDS.
  - *Minimal Missing Values (< 1%)*: Life expectancy, Adult Mortality, BMI, thinness 1-19 years, thinness 5-9 years.
  - *Moderate Missing Values (1-10%)*: Alcohol, Total expenditure, Polio, Diphtheria, Income composition of resources, Schooling.
  - *Significant Missing Values (> 10%)*: Hepatitis B, GDP, Population.
- *Key Observations*:
  - Columns like Hepatitis B, GDP, and Population have over 15 percent missinghissing values.



Missing Values Handling

In this analysis, we are addressing missing values using a two-step approach:

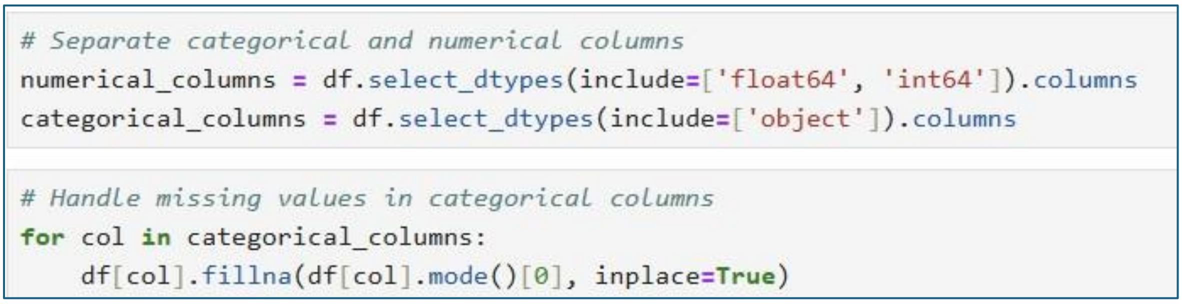
1. **Categorical Columns**:
   - Missing values in categorical columns (e.g., `Country`, `Status`) are imputed using the **mode**, which represents the most frequent value in the column.
   - This method ensures that the imputed values are meaningful and consistent with the existing data distribution.
2. **Numerical Columns**:
   - Missing values in numerical columns are imputed using the **KNNImputer**, which fills missing values based on the mean values of the `k` nearest neighbors in the dataset.
   - The `KNNImputer` is particularly useful for maintaining relationships between features during imputation.



Outlier Treatment

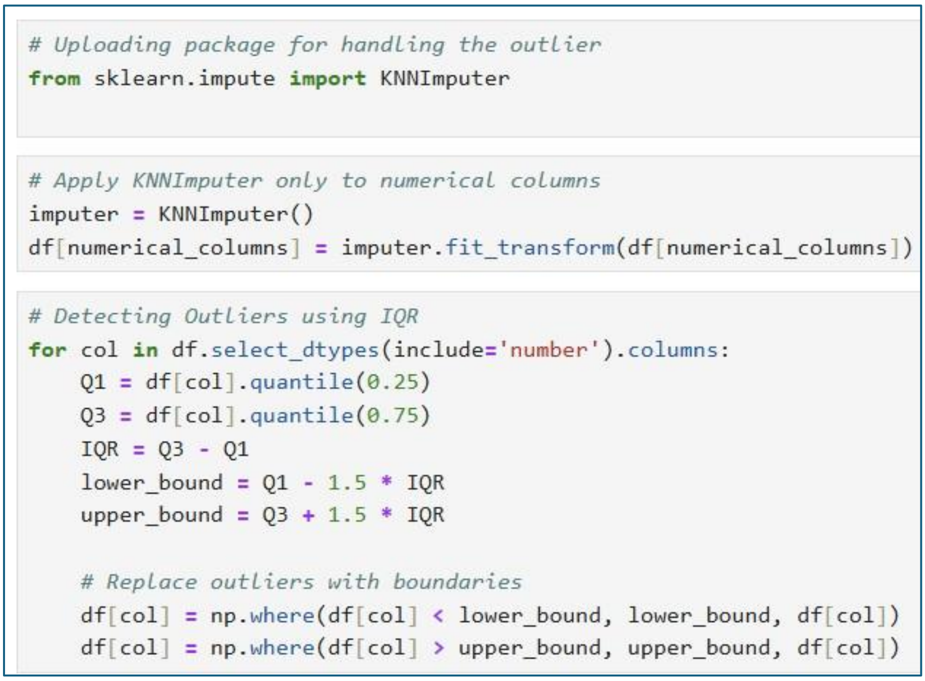
To handle outliers in the dataset, we use the **Interquartile Range (IQR) method**:

1. **Detection**:
   - The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of a feature.
   - Outliers are identified as values lying below `Q1 - 1.5 * IQR` or above `Q3 + 1.5 * IQR`.
2. **Treatment**:
   - Outliers are capped to the lower and upper boundaries (`Q1 - 1.5 * IQR` and `Q3 + 1.5 * IQR`, respectively).
   - This approach retains all data points while reducing the impact of extreme values on the analysis.



```
# Separate categorical and numerical columns
numerical_columns = df.select_dtypes(include=['float64', 'int64']).columns
categorical_columns = df.select_dtypes(include=['object']).columns
```

```
# Handle missing values in categorical columns
for col in categorical_columns:
    df[col].fillna(df[col].mode()[0], inplace=True)
```



```
# Uploading package for handling the outlier
from sklearn.impute import KNNImputer
```

```
# Apply KNNImputer only to numerical columns
imputer = KNNImputer()
df[numerical_columns] = imputer.fit_transform(df[numerical_columns])
```

```
# Detecting Outliers using IQR
for col in df.select_dtypes(include='number').columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Replace outliers with boundaries
    df[col] = np.where(df[col] < lower_bound, lower_bound, df[col])
    df[col] = np.where(df[col] > upper_bound, upper_bound, df[col])
```

7. **Outlier Treatment using IQR:**

   Outliers in numerical columns were detected using the Interquartile Range (IQR) method. Values below Q1 - 1.5*IQR or above Q3 + 1.5*IQR were identified as outliers.

These outliers were capped to the lower and upper boundaries, reducing their impact without discarding any data.

8. **Missing Value Handling:**

   For categorical columns, missing values were filled using the mode (most frequent value), ensuring the imputed values align with the existing data distribution.

   For numerical columns, missing values were handled using the KNNImputer, which imputes missing values based on the nearest neighbors, maintaining relationships between features.

- **Let's understand Missing Value Imputation**

   i. **Categorical Columns:** Missing values in categorical columns (e.g., Country, Status) were imputed using the mode.

   The mode represents the most frequently occurring value in a column, ensuring the imputed values align with the existing data distribution.

   Why Mode?

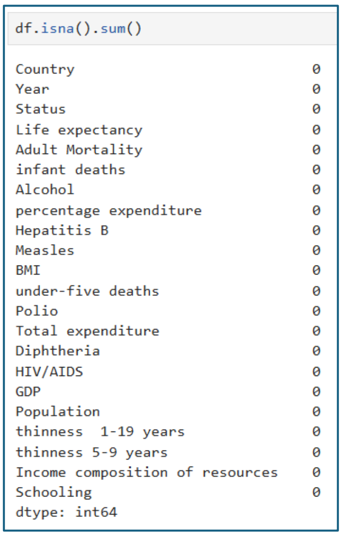
   Mode imputation is simple and effective for categorical data, as it fills missing values with the most likely category without introducing noise.

   ii. **Numerical Columns:** Missing values in numerical columns were imputed using the K-Nearest Neighbors (KNN) Imputer. This method estimates missing values based on the average of the k nearest data points, preserving the relationships and patterns within the data.

Why KNN Imputer?

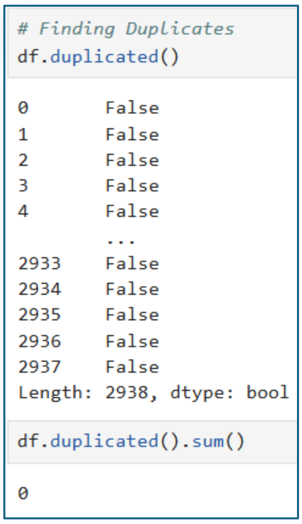
KNN Imputer is particularly useful when numerical features have strong correlations or dependencies. It considers the similarity between data points, resulting in more realistic imputations compared to static methods like mean or median.

Let’s see if we have any null values left in the dataset: We are left with ZERO null values.



```
df.isna().sum()

Country                            0
Year                               0
Status                             0
Life expectancy                    0
Adult Mortality                    0
infant deaths                      0
Alcohol                            0
percentage expenditure             0
Hepatitis B                        0
Measles                            0
BMI                                0
under-five deaths                  0
Polio                              0
Total expenditure                  0
Diphtheria                         0
HIV/AIDS                           0
GDP                                0
Population                         0
thinness  1-19 years               0
thinness 5-9 years                 0
Income composition of resources    0
Schooling                          0
dtype: int64
```

G. **Let’s find if our dataset contains Duplicate values:** Below image shows that our dataset has ZERO duplicate values and now it is suitable for analysis or modeling.



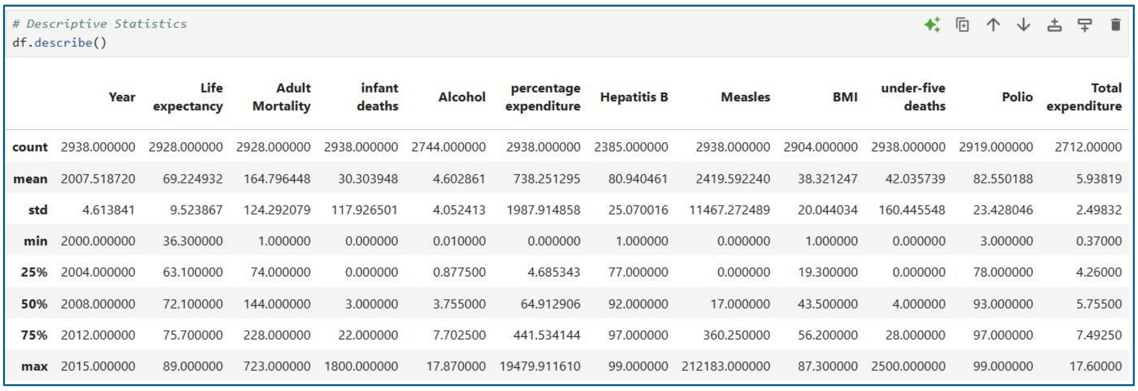
```
# Finding Duplicates
df.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
2933    False
2934    False
2935    False
2936    False
2937    False
Length: 2938, dtype: bool
```

```
df.duplicated().sum()
```

```
0
```

10. **Descriptive Statistics:** df. Describe () is a crucial step for initial data exploration, offering insights into data quality and characteristics. The image below summarizes key metrics like count, mean, standard deviation, and range for numerical features such as life expectancy, adult mortality, and GDP. These statistics help identify data distribution, outliers, and missing values to guide further analysis.

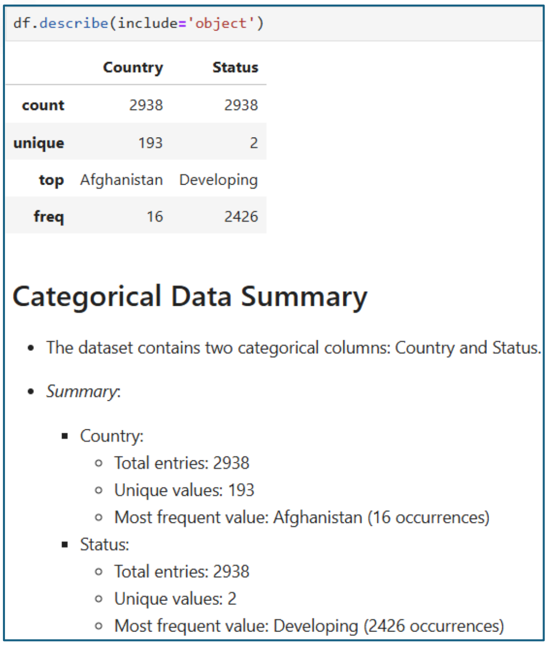


```
# Descriptive Statistics
df.describe()
```

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.000000	2919.000000	2712.00000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.035739	82.550188	5.93819
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.445548	23.428046	2.49832
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000	0.37000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.000000	78.000000	4.26000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.000000	93.000000	5.75500
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.000000	97.000000	7.49250
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000	17.60000

11. **Categorical Analysis** (df.describe(include='object'))

- The dataset includes two **categorical columns**: Country with 193 unique values and Status with 2 categories (Developing and Developed). The most frequent country is Afghanistan, appearing 16 times, and the Developing status dominates with 2426 occurrences. These summaries help understand the distribution of categorical data and guide further analysis.



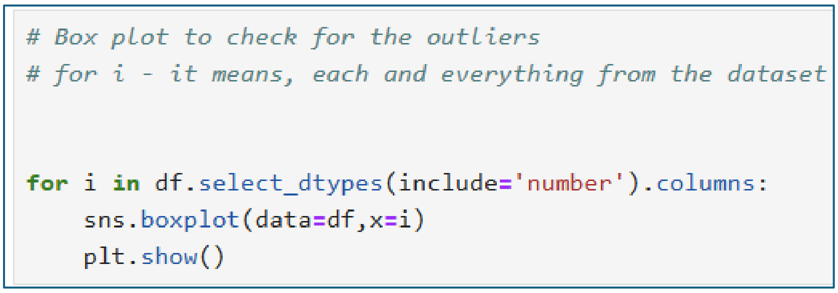
```
df.describe(include='object')
```

	Country	Status
count	2938	2938
unique	193	2
top	Afghanistan	Developing
freq	16	2426

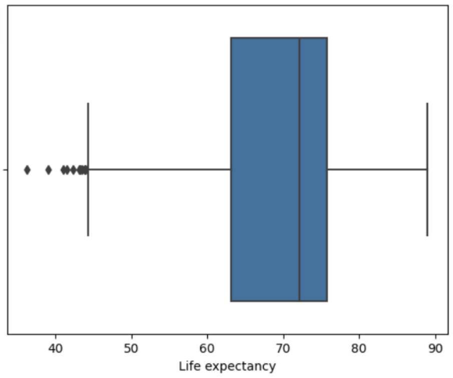
Categorical Data Summary

- The dataset contains two categorical columns: Country and Status.
- *Summary*:
  - Country:
    - Total entries: 2938
    - Unique values: 193
    - Most frequent value: Afghanistan (16 occurrences)
  - Status:
    - Total entries: 2938
    - Unique values: 2
    - Most frequent value: Developing (2426 occurrences)

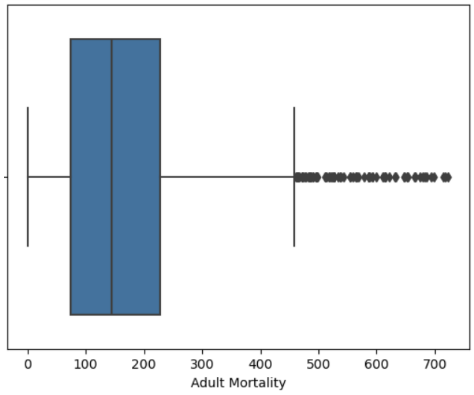
12. Let’s check the **outlier** present in the dataset:



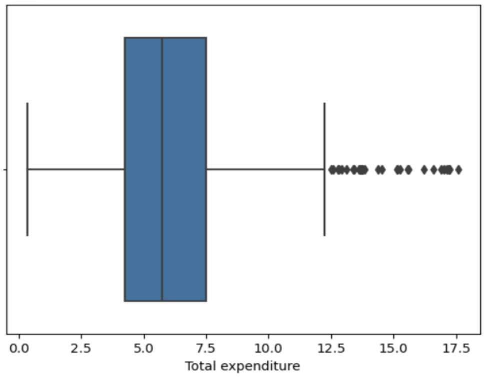
```
# Box plot to check for the outliers
# for i - it means, each and everything from the dataset


for i in df.select_dtypes(include='number').columns:
    sns.boxplot(data=df,x=i)
    plt.show()
```



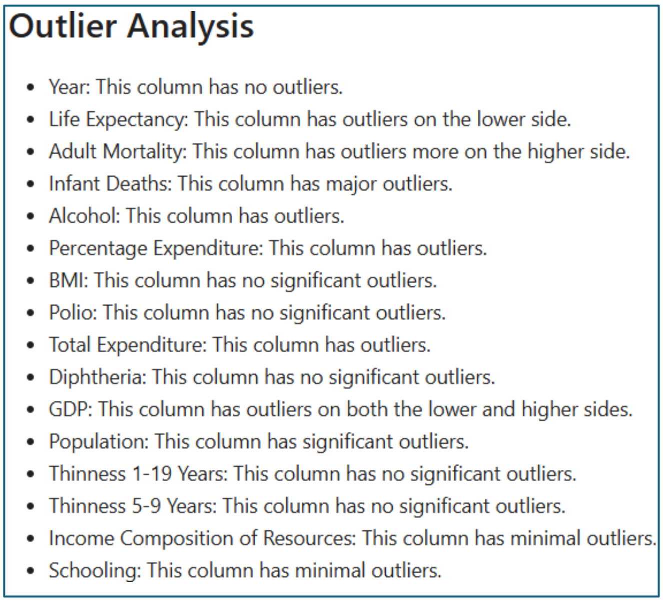
Life Expectancy: Few outliers on the lower end (<40 years), representing countries with significantly low life



Adult Mortality: High outliers (>500 deaths) indicate regions with extreme adult mortality rates.



Total Expenditure: Outliers above 12% represent countries with unusually high healthcare spending.

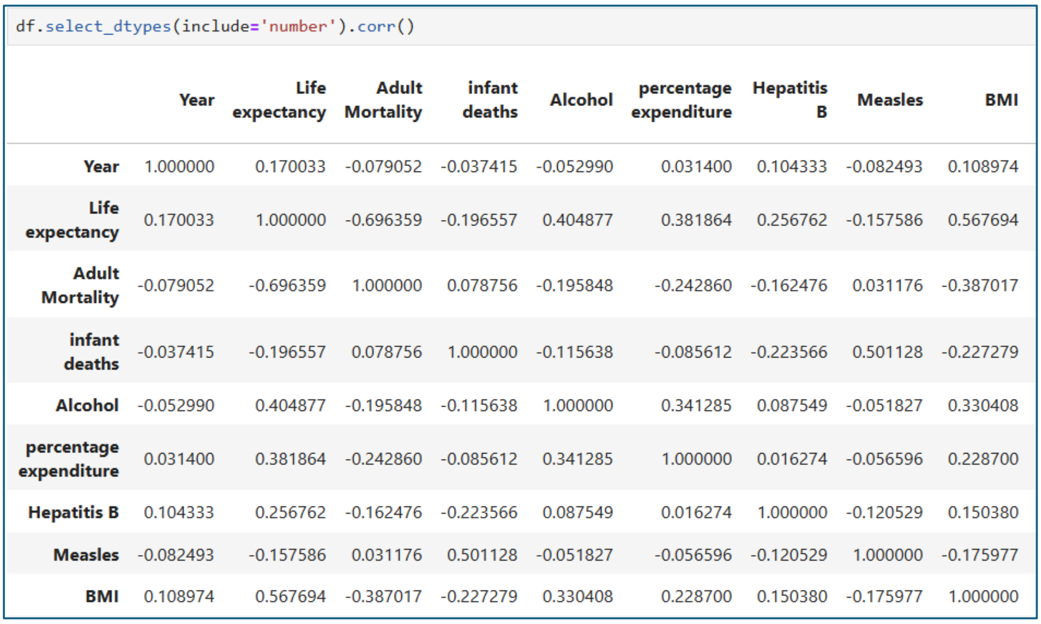


Outlier Analysis

- Year: This column has no outliers.
- Life Expectancy: This column has outliers on the lower side.
- Adult Mortality: This column has outliers more on the higher side.
- Infant Deaths: This column has major outliers.
- Alcohol: This column has outliers.
- Percentage Expenditure: This column has outliers.
- BMI: This column has no significant outliers.
- Polio: This column has no significant outliers.
- Total Expenditure: This column has outliers.
- Diphtheria: This column has no significant outliers.
- GDP: This column has outliers on both the lower and higher sides.
- Population: This column has significant outliers.
- Thinness 1-19 Years: This column has no significant outliers.
- Thinness 5-9 Years: This column has no significant outliers.
- Income Composition of Resources: This column has minimal outliers.
- Schooling: This column has minimal outliers.

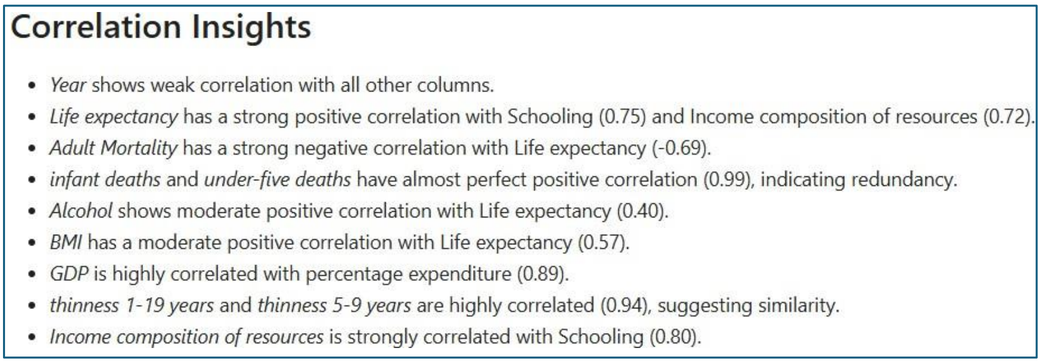
NOTE: The presence of outliers was identified using box plots, highlighting extreme values that may affect analysis and modeling.

13. The **correlation matrix** shows relationships between numerical features, identifying how strongly variables are associated. For example, **life expectancy has a strong negative correlation with adult mortality (-0.6G) and a positive correlation with BMI (0.57).** This analysis helps identify key predictors, redundant variables, and multicollinearity issues, guiding feature selection for modeling.



```
df.select_dtypes(include='number').corr()
```

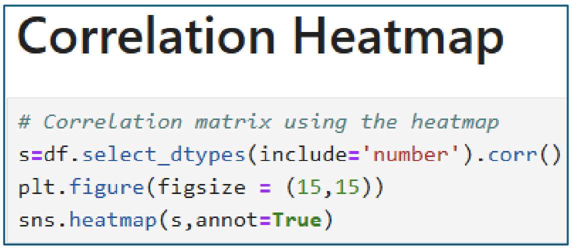
	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI
Year	1.000000	0.170033	-0.079052	-0.037415	-0.052990	0.031400	0.104333	-0.082493	0.108974
Life expectancy	0.170033	1.000000	-0.696359	-0.196557	0.404877	0.381864	0.256762	-0.157586	0.567694
Adult Mortality	-0.079052	-0.696359	1.000000	0.078756	-0.195848	-0.242860	-0.162476	0.031176	-0.387017
infant deaths	-0.037415	-0.196557	0.078756	1.000000	-0.115638	-0.085612	-0.223566	0.501128	-0.227279
Alcohol	-0.052990	0.404877	-0.195848	-0.115638	1.000000	0.341285	0.087549	-0.051827	0.330408
percentage expenditure	0.031400	0.381864	-0.242860	-0.085612	0.341285	1.000000	0.016274	-0.056596	0.228700
Hepatitis B	0.104333	0.256762	-0.162476	-0.223566	0.087549	0.016274	1.000000	-0.120529	0.150380
Measles	-0.082493	-0.157586	0.031176	0.501128	-0.051827	-0.056596	-0.120529	1.000000	-0.175977
BMI	0.108974	0.567694	-0.387017	-0.227279	0.330408	0.228700	0.150380	-0.175977	1.000000



Correlation Insights

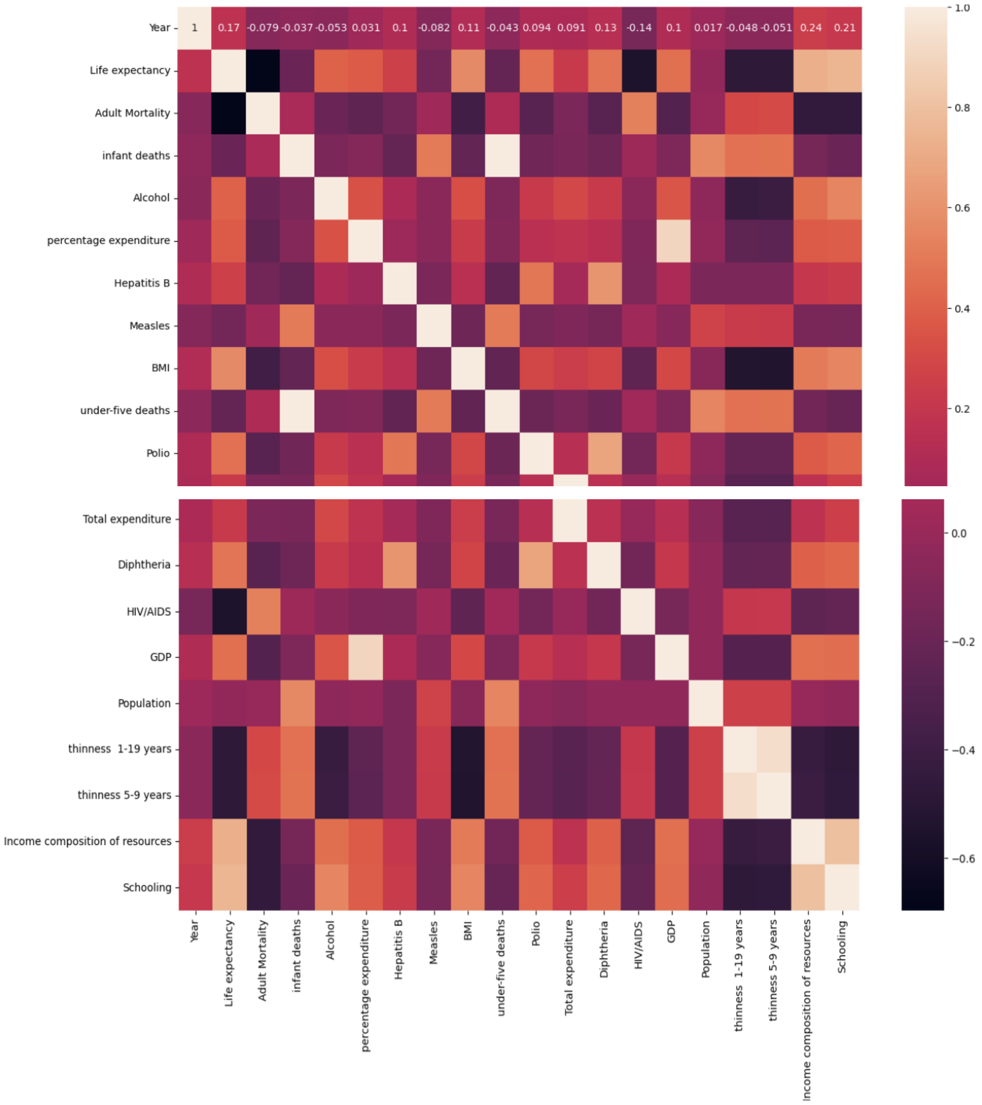
- *Year* shows weak correlation with all other columns.
- *Life expectancy* has a strong positive correlation with Schooling (0.75) and Income composition of resources (0.72).
- *Adult Mortality* has a strong negative correlation with Life expectancy (-0.69).
- *infant deaths* and *under-five deaths* have almost perfect positive correlation (0.99), indicating redundancy.
- *Alcohol* shows moderate positive correlation with Life expectancy (0.40).
- *BMI* has a moderate positive correlation with Life expectancy (0.57).
- *GDP* is highly correlated with percentage expenditure (0.89).
- *thinness 1-19 years* and *thinness 5-9 years* are highly correlated (0.94), suggesting similarity.
- *Income composition of resources* is strongly correlated with Schooling (0.80).

14. The **correlation heatmap** visually represents the strength and direction of relationships between numerical features in the dataset. **It highlights highly correlated variables (e.g., schooling and life expectancy) and helps identify redundant features for better analysis and modeling.**



Correlation Heatmap

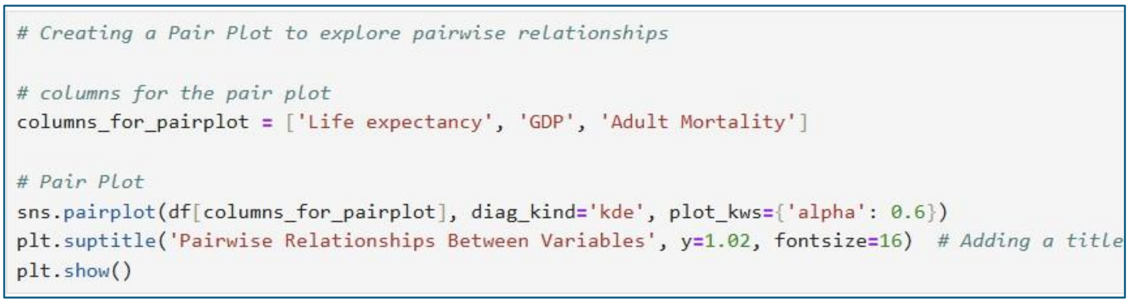
```
# Correlation matrix using the heatmap
s=df.select_dtypes(include='number').corr()
plt.figure(figsize = (15,15))
sns.heatmap(s,annot=True)
```



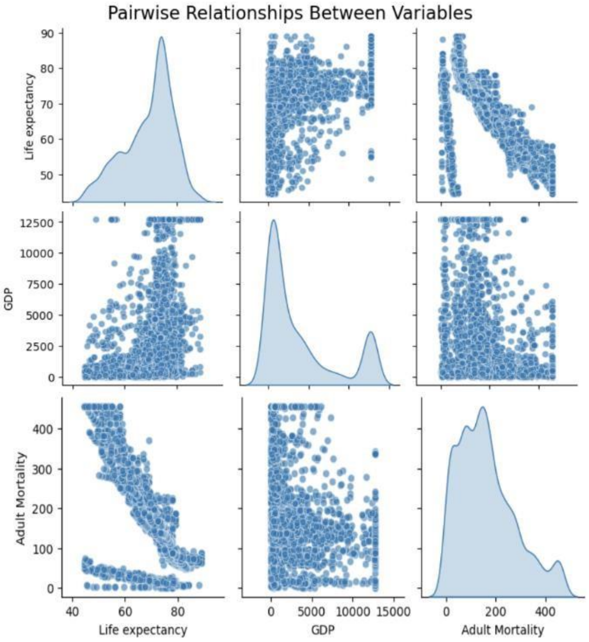
DATA VISUALIZATION

1. **PAIRPLOT: Pairwise relationship between variables.**

   The purpose of this pair plot in the project is to **explore how Life Expectancy, GDP, and Adult Mortality** relate to one another, helping identify patterns and correlations between economic and health factors. **It provides insights into how wealth and mortality impact the overall health outcomes of a population.**



```
# Creating a Pair Plot to explore pairwise relationships

# columns for the pair plot
columns_for_pairplot = ['Life expectancy', 'GDP', 'Adult Mortality']

# Pair Plot
sns.pairplot(df[columns_for_pairplot], diag_kind='kde', plot_kws={'alpha': 0.6})
plt.suptitle('Pairwise Relationships Between Variables', y=1.02, fontsize=16)  # Adding a title
plt.show()
```



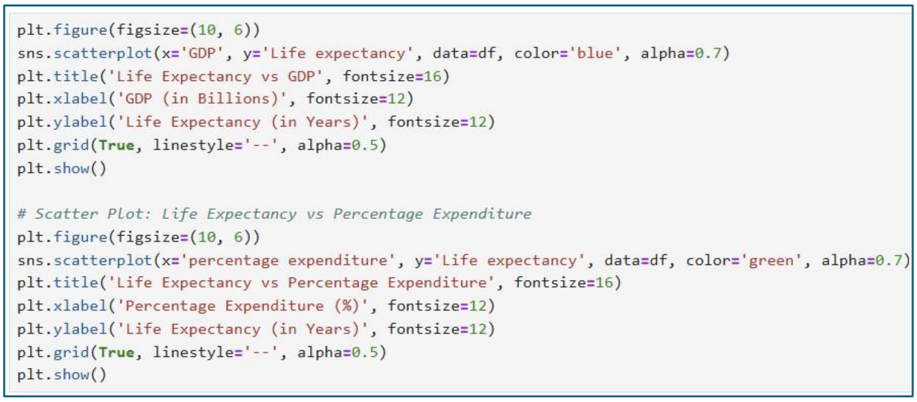
Key Insights:

Economic Impact: GDP is a strong predictor of health outcomes (both life expectancy and adult mortality). Wealthier countries tend to have better healthcare infrastructure, which improves life expectancy and reduces mortality rates.

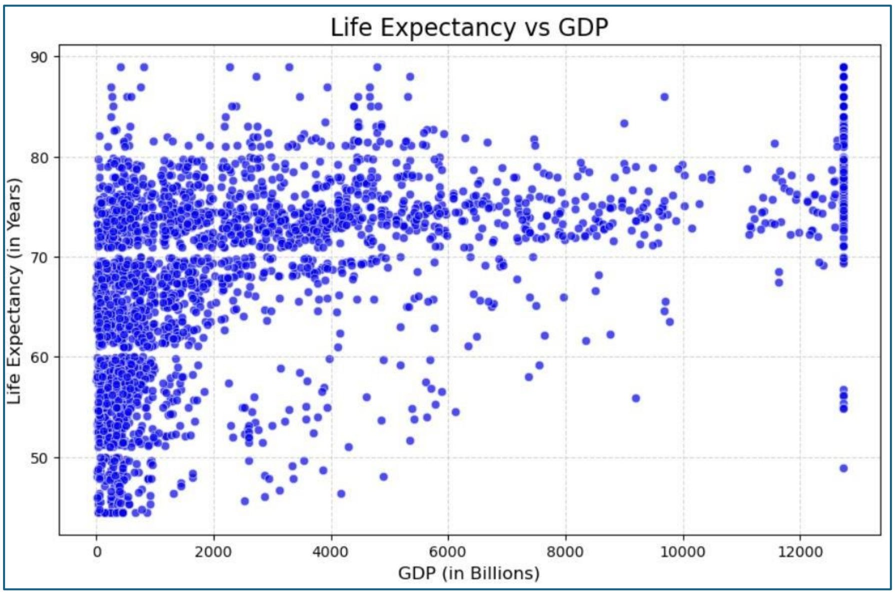
Healthcare Challenges in Low GDP Nations: Countries with low GDP and high adult mortality represent potential areas for focused health and economic development.

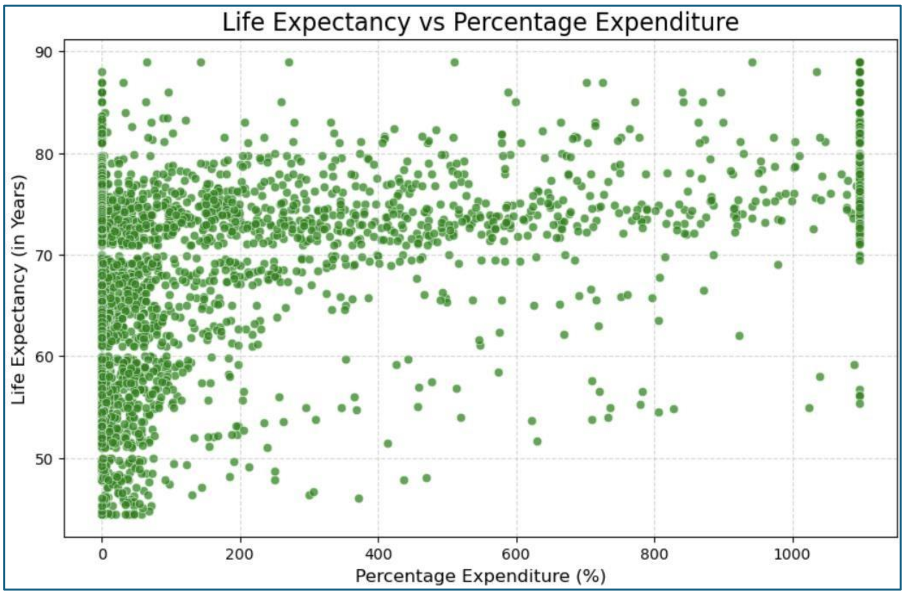
Outliers: There are notable outliers in all relationships, warranting further exploration (e.g., countries with high GDP but lower life expectancy or low GDP and lower-than-expected mortality).

2. **SCATTERPLOT:** The scatter plots visualize relationships between **life expectancy and key factors like GDP and percentage expenditure, making it easier to identify trends, correlations, and outliers.** These visualizations highlight how economic and health spending impact global health outcomes, offering actionable insights for policymakers.



```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='GDP', y='Life expectancy', data=df, color='blue', alpha=0.7)
plt.title('Life Expectancy vs GDP', fontsize=16)
plt.xlabel('GDP (in Billions)', fontsize=12)
plt.ylabel('Life Expectancy (in Years)', fontsize=12)
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()

# Scatter Plot: Life Expectancy vs Percentage Expenditure
plt.figure(figsize=(10, 6))
sns.scatterplot(x='percentage expenditure', y='Life expectancy', data=df, color='green', alpha=0.7)
plt.title('Life Expectancy vs Percentage Expenditure', fontsize=16)
plt.xlabel('Percentage Expenditure (%)', fontsize=12)
plt.ylabel('Life Expectancy (in Years)', fontsize=12)
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```





Explanation of the Scatter Plot: Life Expectancy vs. GDP

This scatter plot illustrates the relationship between Life Expectancy (Y-axis) and GDP (Gross Domestic Product, X-axis) across various countries. Each point represents a country, with its GDP on the horizontal axis and life expectancy on the vertical axis.

- **Key Observations:**
  - **Positive Correlation:** The general trend shows that as GDP increases, life expectancy tends to rise. Wealthier countries (higher GDP) tend to have

longer life expectancy, reflecting better healthcare systems, infrastructure, and living standards.

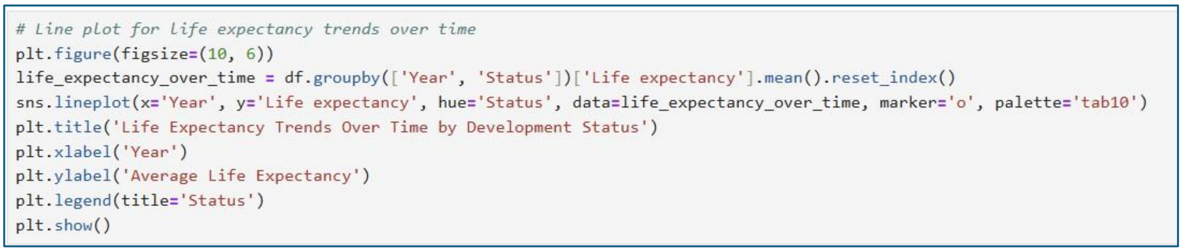
- **Cluster of Low GDP Nations:** A significant cluster of countries is found at the lower GDP range (< 10,000) with varied life expectancies ranging from 40 to 80 years. This indicates a wide disparity in health outcomes among poorer nations, possibly due to differences in healthcare policies or socioeconomic conditions.

- **Explanation of the Scatter Plot: Life Expectancy vs Percentage Expenditure**

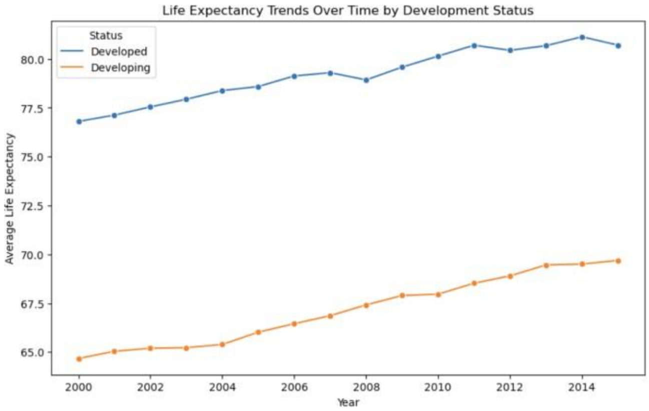
  This scatter plot explores the relationship between Life Expectancy (Y-axis) and Percentage Expenditure on Healthcare (X-axis) across various countries. Each point represents a country, with its healthcare expenditure as a percentage of GDP on the horizontal axis and life expectancy on the vertical axis.

- **Key Observations:**
  - **Clustered Data at Low Expenditure:** Most countries have low healthcare expenditure as a percentage of GDP (< 2,500%). These countries exhibit a wide range of life expectancy, spanning from around 40 to 85 years.
  - **Positive Correlation at Higher Expenditures:** At moderate-to-high healthcare expenditure (between 2,500% and 10,000%), life expectancy tends to stabilize around 70-85 years, showing a generally positive relationship.
  - **Low Expenditure with High Life Expectancy:** Some countries achieve high life expectancy (> 70 years) despite low healthcare expenditure.

3. **LINEPLOT:** The line plot showcases changes in average life expectancy over the years, comparing developed and developing countries. This visualization highlights how **development status influences health outcomes, providing actionable insights for policymakers to address disparities and improve global life expectancy.**



```
# Line plot for life expectancy trends over time
plt.figure(figsize=(10, 6))
life_expectancy_over_time = df.groupby(['Year', 'Status'])['Life expectancy'].mean().reset_index()
sns.lineplot(x='Year', y='Life expectancy', hue='Status', data=life_expectancy_over_time, marker='o', palette='tab10')
plt.title('Life Expectancy Trends Over Time by Development Status')
plt.xlabel('Year')
plt.ylabel('Average Life Expectancy')
plt.legend(title='Status')
plt.show()
```



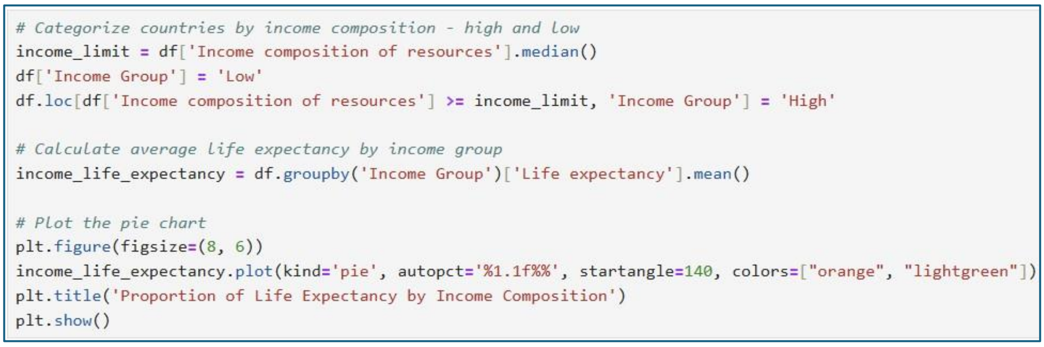
- **Explanation of the Line Plot: Life Expectancy vs. Year**

  This line plot shows the trend in average life expectancy (Y-axis) over time (X-axis) for developed and developing countries. The two lines represent the differences between these groups, highlighting how life expectancy has improved over the years and how development status impacts health outcomes.

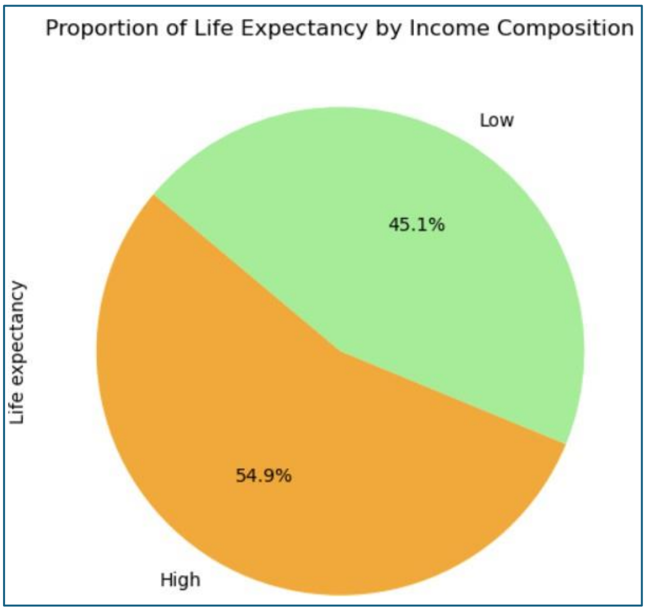
- **Key Observations:**

- **Higher Life Expectancy in Developed Countries:** The blue line (developed countries) consistently shows a higher life expectancy compared to the orange line (developing countries) across all years.
- **Steady Growth in Both Groups:** Both lines demonstrate an upward trend, indicating that life expectancy has been improving over time for both developed and developing countries.
- **Slower Progress in Developing Countries:** The orange line for developing countries shows a slower rate of increase compared to the blue line.

4. **PIECHART:** This pie chart shows how life expectancy is distributed between high and low-income countries. It helps visualize the impact of income on life expectancy, highlighting differences between these groups.



```
# Categorize countries by income composition - high and low
income_limit = df['Income composition of resources'].median()
df['Income Group'] = 'Low'
df.loc[df['Income composition of resources'] >= income_limit, 'Income Group'] = 'High'

# Calculate average life expectancy by income group
income_life_expectancy = df.groupby('Income Group')['Life expectancy'].mean()

# Plot the pie chart
plt.figure(figsize=(8, 6))
income_life_expectancy.plot(kind='pie', autopct='%1.1f%%', startangle=140, colors=["orange", "lightgreen"])
plt.title('Proportion of Life Expectancy by Income Composition')
plt.show()
```

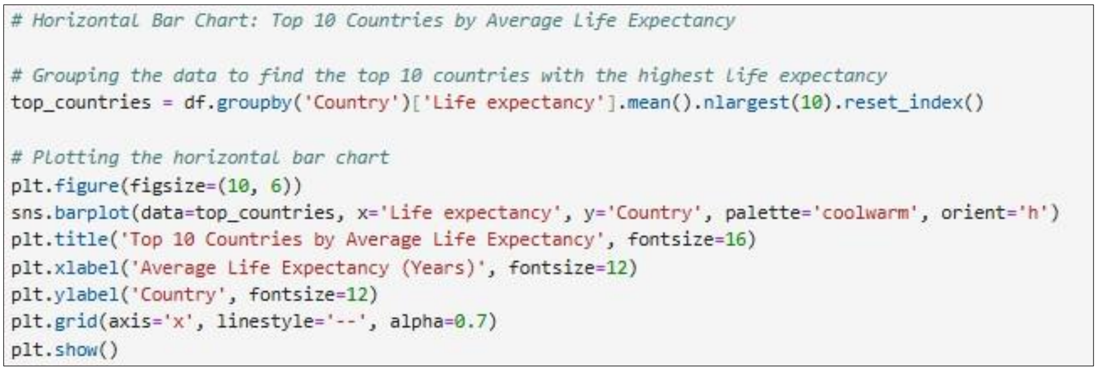


- **Explanation of the Pie Chart**

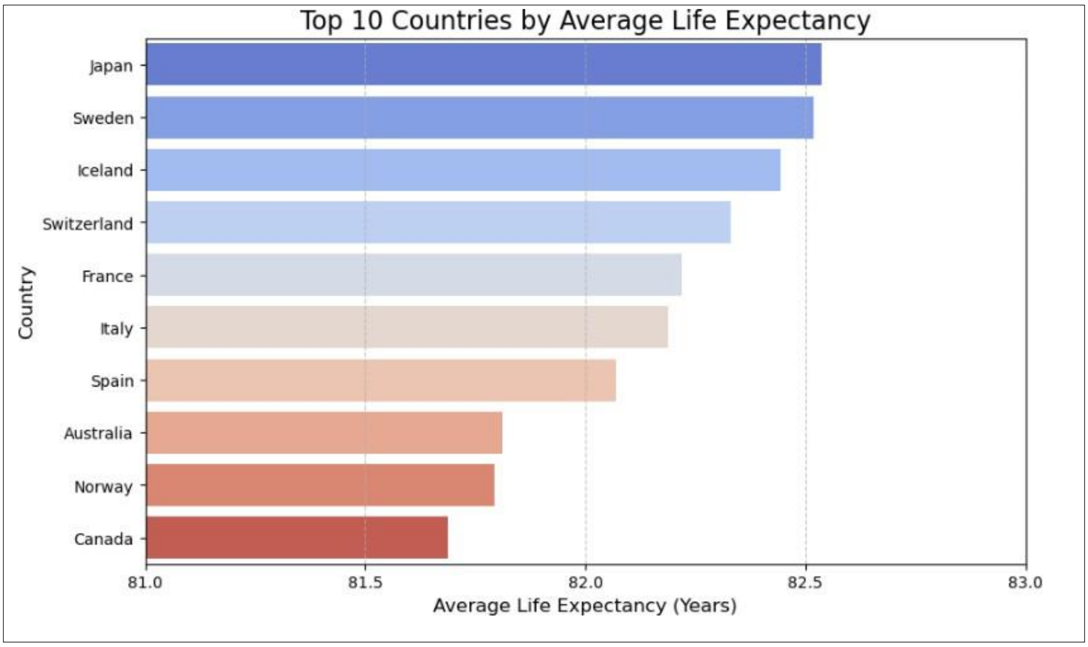
  The pie chart shows the distribution of life expectancy between high-income and low-income countries, with green representing low-income countries and orange representing high-income countries.

  Higher Life Expectancy in High-Income Countries: The orange section (high-income) represents a larger proportion of life expectancy with 54.5%, indicating better health outcomes in wealthier nations.

5. **Horizontal Chart:** This chart aims to identify and highlight the top 10 countries in terms of life expectancy, showcasing global leaders in health outcomes.



```
# Horizontal Bar Chart: Top 10 Countries by Average Life Expectancy

# Grouping the data to find the top 10 countries with the highest life expectancy
top_countries = df.groupby('Country')['Life expectancy'].mean().nlargest(10).reset_index()

# Plotting the horizontal bar chart
plt.figure(figsize=(10, 6))
sns.barplot(data=top_countries, x='Life expectancy', y='Country', palette='coolwarm', orient='h')
plt.title('Top 10 Countries by Average Life Expectancy', fontsize=16)
plt.xlabel('Average Life Expectancy (Years)', fontsize=12)
plt.ylabel('Country', fontsize=12)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
```



- **Explanation of the Horizontal Chart:**

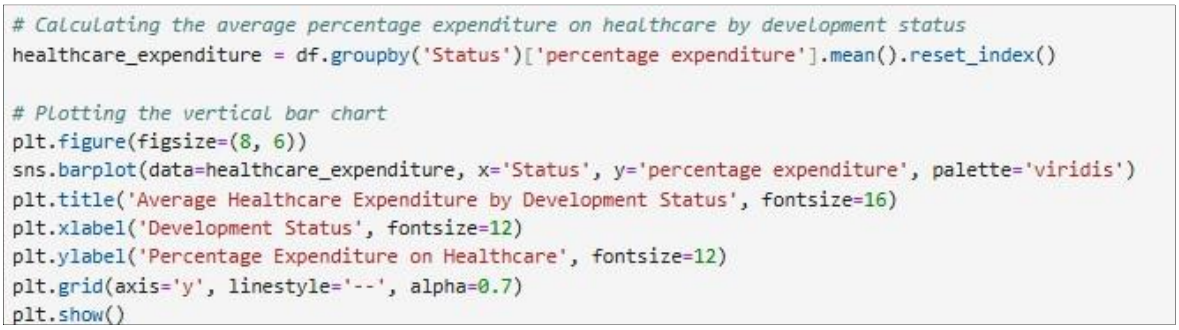
  This horizontal bar chart visualizes the top 10 countries with the highest average life expectancy, showcasing which nations lead in global health outcomes.

- **Key Observations:**
  - **Japan Leads in Life Expectancy:** Japan has the highest life expectancy among all countries, closely followed by Sweden and Iceland.

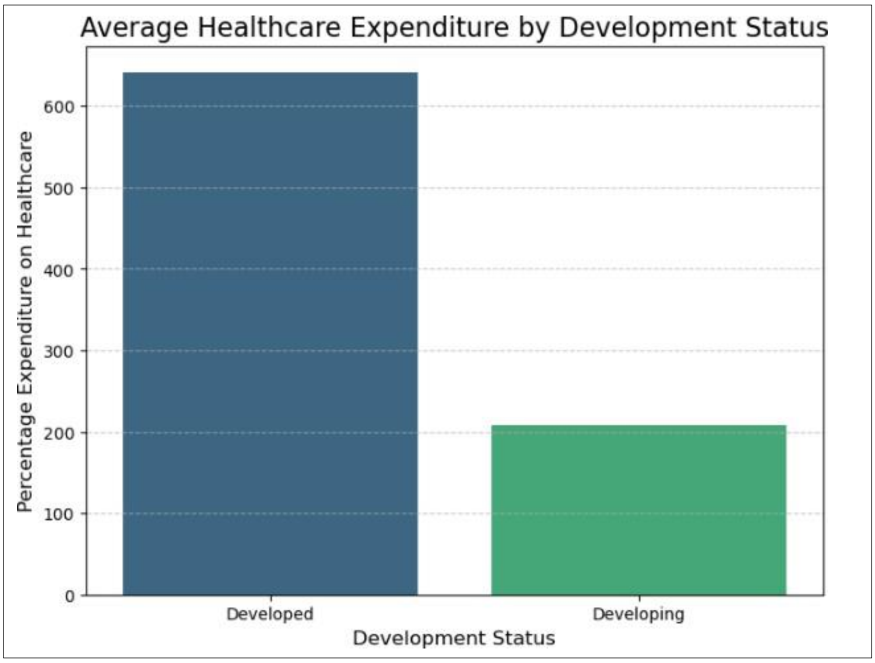
- **Top-Performing Countries Are Developed:** All top-ranking countries are developed nations, reflecting their robust healthcare systems and high standards of living.

- **Insights for Policymakers:**
  - **Study Best Practices in Leading Countries:** Developing countries can learn from the policies and practices of these top-performing nations to improve their healthcare systems.
  - **Sustain High Standards:** Policymakers in these countries should continue investing in healthcare to maintain their global leadership in life expectancy.

6. **Vertical Bar Chart**: The chart compares **Healthcare Expenditure as a percentage of GDP** between developed and developing countries, highlighting disparities in healthcare investment.



```
# Calculating the average percentage expenditure on healthcare by development status
healthcare_expenditure = df.groupby('Status')['percentage expenditure'].mean().reset_index()

# Plotting the vertical bar chart
plt.figure(figsize=(8, 6))
sns.barplot(data=healthcare_expenditure, x='Status', y='percentage expenditure', palette='viridis')
plt.title('Average Healthcare Expenditure by Development Status', fontsize=16)
plt.xlabel('Development Status', fontsize=12)
plt.ylabel('Percentage Expenditure on Healthcare', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



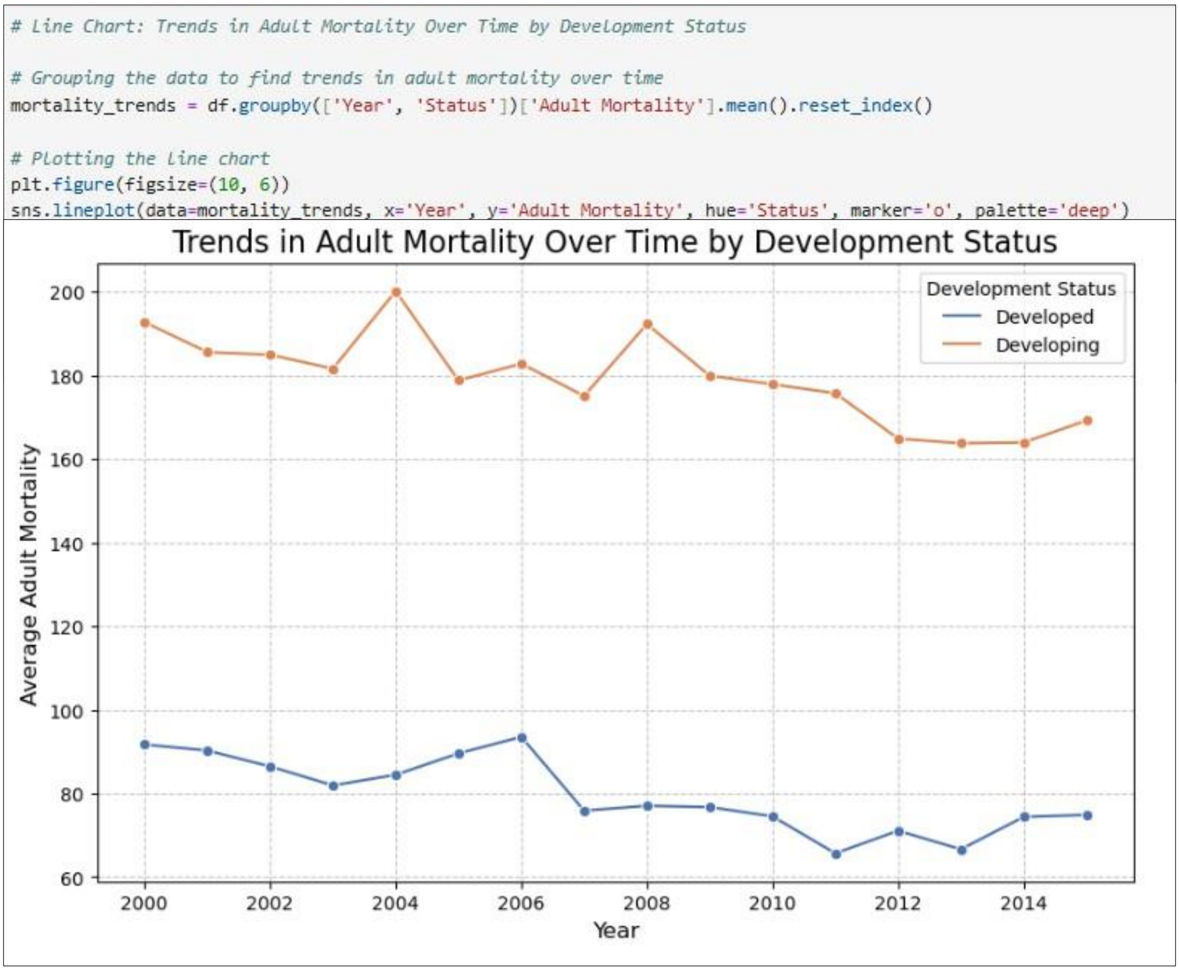
- **Explanation of the Vertical Bar Chart:**

  The vertical bar chart shows the average percentage of GDP spent on healthcare in developed and developing countries. It provides a clear comparison of healthcare investment by development status.

- **Key Observations:**
  - **Higher Healthcare Spending in Developed Countries:** Developed countries allocate a significantly larger percentage of their GDP to healthcare compared to developing countries.
  - **Developing Countries Lag Behind:** Developing nations spend much less on healthcare, which can hinder their ability to improve public health outcomes.

- **Insights for Policymakers:**
  - **Increase Investment in Healthcare:** Developing countries should prioritize healthcare funding to improve life expectancy and reduce mortality rates.
  - **Efficient Use of Resources in Developed Countries:** Developed countries should focus on optimizing healthcare expenditure to ensure long-term sustainability.

7. **Line Chart:** The chart visualizes **Trends in Adult Mortality Rates** (deaths per 1,000 adults) over time for developed and developing countries, showcasing progress or disparities in reducing mortality.



```
# Line Chart: Trends in Adult Mortality Over Time by Development Status

# Grouping the data to find trends in adult mortality over time
mortality_trends = df.groupby(['Year', 'Status'])['Adult Mortality'].mean().reset_index()

# Plotting the line chart
plt.figure(figsize=(10, 6))
sns.lineplot(data=mortality_trends, x='Year', y='Adult Mortality', hue='Status', marker='o', palette='deep')
```

- **Explanation of the Line Chart:**

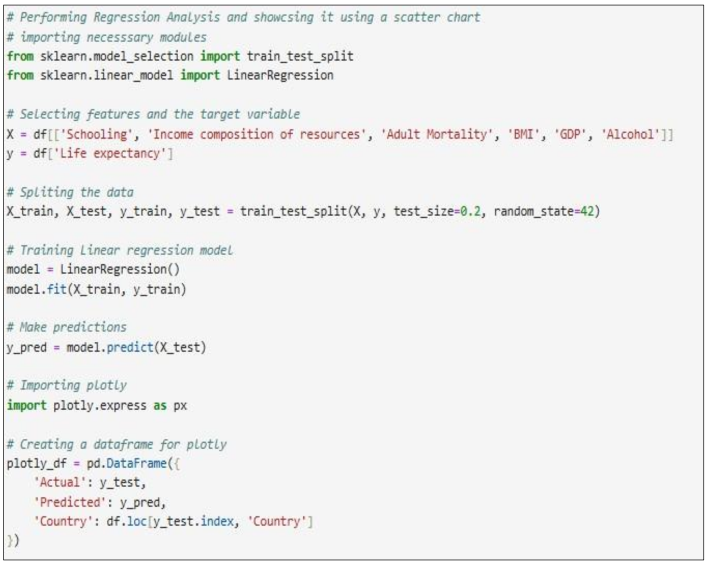
  The line chart shows the average adult mortality rates for developed and developing countries over several years. Each line represents the trend for a specific development status, with markers to highlight key data points.

- **Key Observations:**
  - **Higher Mortality in Developing Countries:** Developing countries have consistently higher adult mortality rates compared to developed countries.
  - **Declining Trend in Mortality Rates:** Both developed and developing countries show a downward trend, indicating global improvements in healthcare and living conditions.

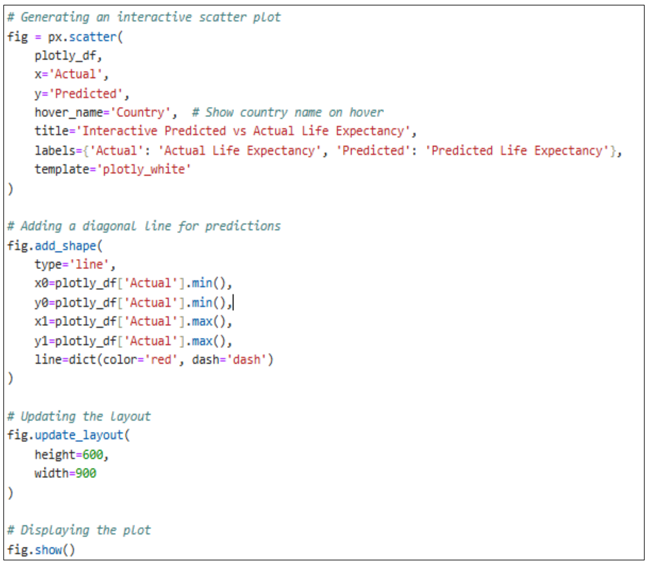
- **Insights for Policymakers:**
  - **Focus on Reducing Mortality in Developing Countries:** Policymakers should prioritize healthcare interventions in developing countries to accelerate the decline in mortality rates.
  - **Maintain Progress in Developed Countries:** Developed nations should continue investing in preventive care and managing chronic diseases to sustain their low mortality rates.

PREDICTIVE MODELS

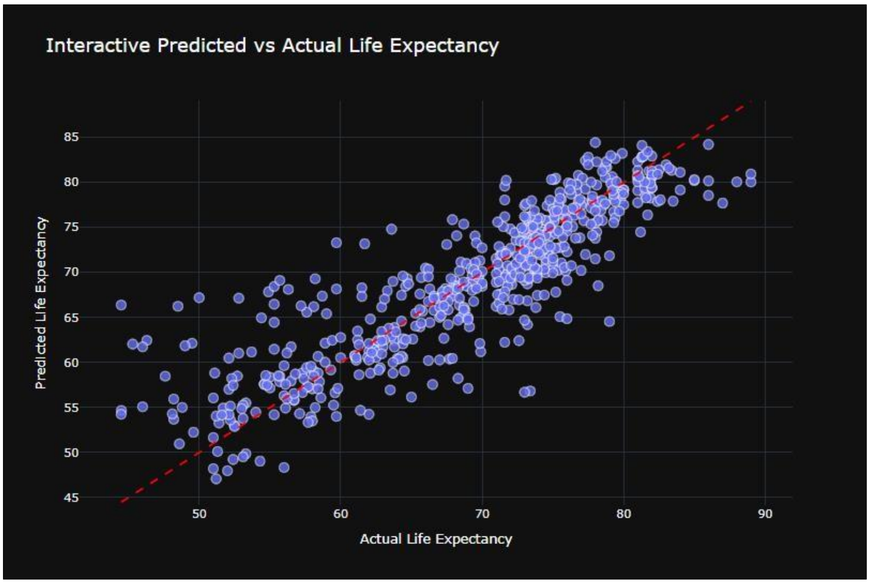
Linear Regression on a scatter plot: The scatterplot compares actual vs. predicted life expectancy values. The red diagonal line represents perfect predictions. Points close to the line indicate accurate predictions, while deviations show prediction errors.



```
# Performing Regression Analysis and showcsing it using a scatter chart
# importing necesssary modules
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Selecting features and the target variable
X = df[['Schooling', 'Income composition of resources', 'Adult Mortality', 'BMI', 'GDP', 'Alcohol']]
y = df['Life expectancy']

# Spliting the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Training Linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Importing plotly
import plotly.express as px

# Creating a dataframe for plotly
plotly_df = pd.DataFrame({
    'Actual': y_test,
    'Predicted': y_pred,
    'Country': df.loc[y_test.index, 'Country']
})
```



```
# Generating an interactive scatter plot
fig = px.scatter(
    plotly_df,
    x='Actual',
    y='Predicted',
    hover_name='Country',  # Show country name on hover
    title='Interactive Predicted vs Actual Life Expectancy',
    labels={'Actual': 'Actual Life Expectancy', 'Predicted': 'Predicted Life Expectancy'},
    template='plotly_white'
)

# Adding a diagonal line for predictions
fig.add_shape(
    type='line',
    x0=plotly_df['Actual'].min(),
    y0=plotly_df['Actual'].min(),
    x1=plotly_df['Actual'].max(),
    y1=plotly_df['Actual'].max(),
    line=dict(color='red', dash='dash')
)

# Updating the layout
fig.update_layout(
    height=600,
    width=900
)

# Displaying the plot
fig.show()
```

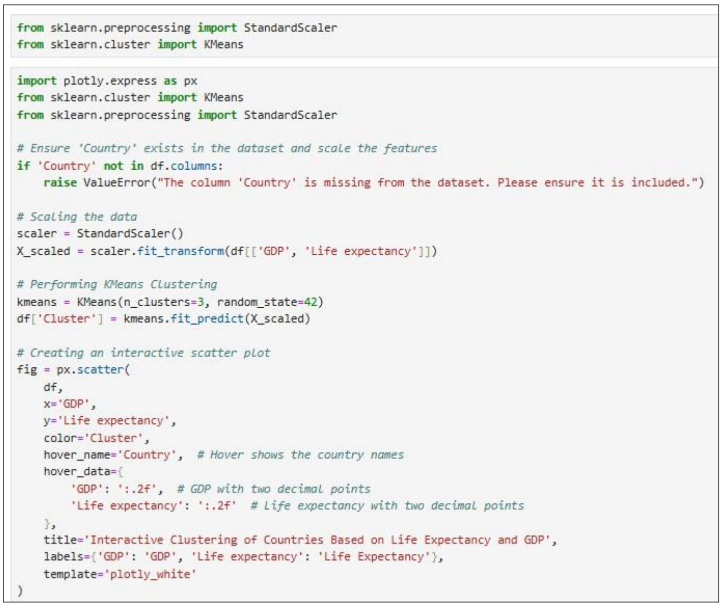


- **Explanation of the scatter chart:**

The scatter chart visualizes the relationship between actual and predicted life expectancy, allowing you to hover over each dot to show individual country information.

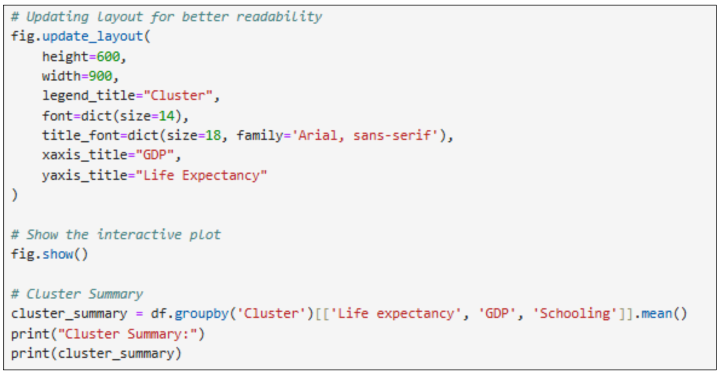
- **Key Patterns:**
    - **Strong Positive Correlation:** Most points align with the red diagonal, showing the model predicts life expectancy accurately for most countries.
    - **Clustered Predictions:** Predictions are densest in the 60-80 life expectancy range, with smaller errors compared to lower ranges.
    - **Outliers:** Some points deviate significantly, indicating countries with unique factors not fully captured by the model.
    - **Feature Importance:** Economic resources, education, and reduced adult mortality strongly influence life expectancy.

Cluster Chart: The chart visualizes clusters of countries based on GDP and life expectancy, highlighting groups with similar economic and health profiles for targeted policy insights

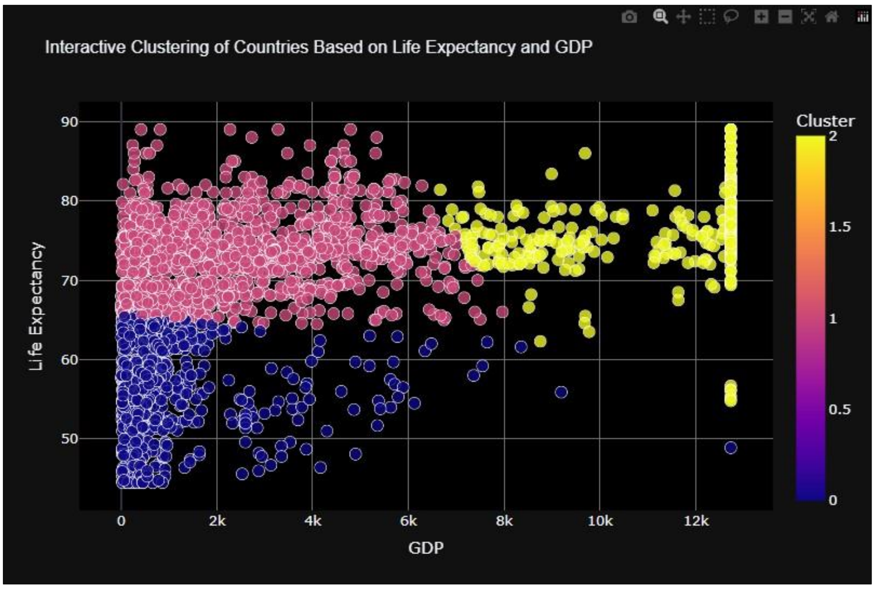


```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

```python
import plotly.express as px
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Ensure 'Country' exists in the dataset and scale the features
if 'Country' not in df.columns:
    raise ValueError("The column 'Country' is missing from the dataset. Please ensure it is included.")

# Scaling the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df[['GDP', 'Life expectancy']])

# Performing KMeans Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(X_scaled)

# Creating an interactive scatter plot
fig = px.scatter(
    df,
    x='GDP',
    y='Life expectancy',
    color='Cluster',
    hover_name='Country',  # Hover shows the country names
    hover_data={
        'GDP': ':.2f',  # GDP with two decimal points
        'Life expectancy': ':.2f'  # Life expectancy with two decimal points
    },
    title='Interactive Clustering of Countries Based on Life Expectancy and GDP',
    labels={'GDP': 'GDP', 'Life expectancy': 'Life Expectancy'},
    template='plotly_white'
)
```



```python
# Updating layout for better readability
fig.update_layout(
    height=600,
    width=900,
    legend_title="Cluster",
    font=dict(size=14),
    title_font=dict(size=18, family='Arial, sans-serif'),
    xaxis_title="GDP",
    yaxis_title="Life Expectancy"
)

# Show the interactive plot
fig.show()

# Cluster Summary
cluster_summary = df.groupby('Cluster')[['Life expectancy', 'GDP', 'Schooling']].mean()
print("Cluster Summary:")
print(cluster_summary)
```



- **Explanation of the Chart:**

  The chart shows clusters of countries based on their GDP (x-axis) and Life Expectancy (y-axis). Each dot represents a country, and the colors indicate different clusters formed by grouping countries with similar characteristics.

This helps us visually understand patterns and differences among countries in terms of economic development and health outcomes.

- **Key Patterns:**
  - **Relationship Between GDP and Life Expectancy:**
    - A positive trend is visible: countries with higher GDP tend to have longer life expectancy. This indicates that economic growth supports better healthcare and living standards.
  - **Clusters Highlight Outliers:**
    - The clusters help identify countries that deviate from the trend, such as countries with relatively high GDP but lower life expectancy, or vice versa.

INSIGHTS DRAWN BASED ON OUR MODELS

- **Key Observations (Scatter Chart - Linear Regression):**

  1. **Model Performance:** The R² Score (0.73) indicates that the model explains about 73% of the variance within life expectancy, which is quite strong. The Root Mean Squared Error (RMSE) of 4.81 tells us that the average prediction error is around 4.81 years, which is withing the margin of error for our analysis.
  2. **Predicted vs. Actual Trend:** We see the points generally align along the red diagonal line, showing that the model predicts life expectancy well for most countries. However, there are some outliers (points far from the line) where the predicted life expectancy deviates significantly from the actual value. These outliers could indicate countries with unique conditions which might need to look into.
  3. **Feature Importance:** The income composition of resources has the highest positive coefficient (9.16), indicating its strong influence on life expectancy. Schooling is the second most impactful factor, telling us how important of a role education has in improving health outcomes. Adult Mortality has a negative coefficient (-0.03), highlighting its adverse effect on life expectancy which makes sense.
  4. **Cluster of Points:** A dense cluster of points around the 60-80 life expectancy range suggests that most countries fall within this range, with relatively small prediction errors. The spread does widen at the lower end (<60 years), indicating greater variability in predicting life expectancy for these countries.

- **Insights for Policymakers (Scatter Chart - Linear Regression):**

1. **Invest in Education:** The strong positive impact of schooling underscores the importance of improving access to quality education to enhance life expectancy.
2. **Boost Economic Resources:** The high coefficient for income composition suggests that economic stability and equitable resource distribution are critical for better health outcomes.
3. **Address Mortality Rates:** The negative correlation with adult mortality highlights the need for policies aimed at reducing premature deaths, such as improving healthcare access and combating diseases.
4. **Examine Outliers**: Countries with significant prediction errors (outliers) may require tailored solutions. Investigate these cases to identify factors not captured in the model, such as political stability, environmental conditions, or cultural influences.
5. **Focus on Vulnerable Regions:** The greater variability at lower life expectancy levels indicates that underdeveloped or developing nations may benefit most from targeted interventions in healthcare, education, and economic development.

- **Key Observations (Cluster Chart):**
  1. **Cluster 0 (Purple):**
     - Countries in this group have the lowest GDP and shortest life expectancy.
     - These countries also tend to have lower levels of education and health facilities.

   - Likely represents underdeveloped nations facing challenges like poverty, limited access to healthcare, and poor infrastructure.

2. **Cluster 1 (Teal):**
   a. Countries in this group have the highest GDP and longest life expectancy.
   b. They also enjoy higher education levels and access to advanced healthcare systems.
   c. Likely corresponds to developed nations with strong economies and excellent public health standards.
3. **Cluster 2 (Yellow):**
   a. This group represents countries with mid-range GDP and moderate life expectancy.
   b. These nations are on the path to growth, with improving education systems and healthcare.
   c. Likely represents developing countries, transitioning towards better economic and health outcomes.

- **Insights for Policymakers (Cluster Chart):**
  1. **Cluster 0 (Underdeveloped Nations):**
     - **Needs:** Basic healthcare, improved education, and poverty alleviation.
     - **Actions:** International aid, healthcare access programs, and investments in education and infrastructure.
  2. **Cluster 1 (Developed Nations):**

- **Needs:** Address aging populations and preventive healthcare to sustain high life expectancy.
- **Actions:** Invest in healthcare innovation, support for elderly care, and maintain economic stability.

3. **Cluster 2 (Developing Nations):**
    - **Needs:** Continued economic growth and improved health infrastructure.
    - **Actions:** Strengthen higher education systems, expand healthcare access, and create policies to support sustainable development.

CONCLUSION

The analysis of factors influencing global life expectancy has revealed critical insights that can guide policymakers and public health organizations. The robust data preparation, visualization, and predictive modeling have highlighted key determinants such as GDP, education, and healthcare spending. These factors exhibit strong correlations with life expectancy, emphasizing the interplay between economic stability, educational access, and healthcare quality in shaping health outcomes.

Key findings include:

1. **Economic and Educational Impact**: Countries with higher GDP and improved educational systems consistently show better health outcomes, underlining the importance of economic growth and access to quality education.
2. **Healthcare Investment**: Efficient allocation of resources to healthcare expenditure positively correlates with increased life expectancy, even in low-income nations.
3. **Targeted Interventions for Vulnerable Populations**: Developing and underdeveloped regions exhibit significant variability in life expectancy, highlighting the need for tailored interventions addressing poverty, healthcare access, and education.

The predictive models employed demonstrate strong performance, with meaningful insights into feature importance. Notable outliers suggest that additional socio-political and environmental factors could further enrich the analysis.

This study advocates for collaborative efforts among global stakeholders to address disparities in health outcomes. Investing in education, economic stability, and equitable healthcare access remains pivotal for enhancing life expectancy, particularly in underdeveloped and developing regions. Future research could explore incorporating additional dimensions such as environmental sustainability and cultural influences to provide a more holistic understanding of health outcome