# Data Science

# Mini Project
# On
# Life Expectancy Analysis

By
Pratyush Mishra

# INDEX

1. Introduction

2. Objective

3. Tools and Libraries Required

4. Dataset and Preprocessing

5. Model Architecture and Training

6. Model Evaluation

7. Visualizations, Insights, and Observations

8. Source Code and Outputs

9. Conclusion

# INTRODUCTION

Life expectancy is one of the most important metrics used to assess the health status and quality of life within populations. It provides an estimate of the average number of years a person is expected to live, assuming that current mortality rates remain constant throughout their lifetime. This indicator is influenced by a wide range of factors including healthcare quality, socio-economic conditions, nutrition, education, disease prevalence, and environmental conditions.

Analyzing life expectancy across countries and over time allows researchers to understand the underlying determinants that contribute to longevity and to identify disparities between different regions and populations. With growing global attention on sustainable development and public health, understanding these factors is critical for guiding health policies and improving outcomes.

In this project, we utilize a global life expectancy dataset which includes variables such as adult mortality rates, immunization coverage, income composition of resources, schooling years, and other health-related indicators. Through rigorous data preprocessing, exploratory data analysis, and statistical modeling including correlation analysis and multivariate linear regression, we explore the relationships between these factors and life expectancy.

Our goal is to uncover significant predictors of life expectancy and to quantify their effects using statistical methods. These insights can support policymakers, public health officials, and researchers in designing targeted interventions aimed at increasing life expectancy and reducing health inequities worldwide.

# OBJECTIVE

The primary objective of this project is to analyze and understand the key factors influencing life expectancy across different countries using statistical and machine learning techniques. Specifically, the project aims to:

- Explore the relationships between life expectancy and various socio-economic, demographic, and health-related variables.
- Identify the most significant predictors of life expectancy through correlation analysis and hypothesis testing.
- Develop a multivariate linear regression model to quantify the impact of multiple factors on life expectancy simultaneously.
- Evaluate the statistical significance and strength of these predictors to derive meaningful insights.
- Visualize data distributions, relationships, and model results to aid interpretation.
- Provide actionable conclusions and recommendations that can inform public health policy and resource allocation.

Ultimately, this project seeks to contribute to the understanding of life expectancy determinants, enabling better strategies for improving population health and longevity.

# TOOLS AND LIBRARIES USED

This project was implemented using the Python programming language, leveraging its powerful data analysis and visualization libraries. The following tools and libraries were utilized:

- Python 3.x: The primary programming language used for data processing, analysis, and modeling.
- Pandas: For efficient data manipulation, cleaning, and preprocessing of the dataset.
- NumPy: To perform numerical operations and support array computations.
- Matplotlib: A fundamental plotting library used for creating static, animated, and interactive visualizations.
- Seaborn: Built on Matplotlib, Seaborn provides a higher-level interface for drawing attractive and informative statistical graphics.
- SciPy: Used for performing advanced statistical tests such as Pearson correlation and t-tests.
- Statsmodels: A library for estimating and interpreting statistical models, specifically used here for linear regression analysis.
- Jupyter Notebook (optional): An interactive computing environment used to write and run code, visualize outputs, and document the workflow.

These tools collectively facilitated comprehensive exploratory data analysis, hypothesis testing, regression modeling, and insightful visualizations in a reproducible manner.

# DATASET AND PREPROCESSING

The dataset utilized in this project is the Life Expectancy Data spanning from 2000 to 2015 for 193 countries. It encompasses various health, economic, and demographic indicators that potentially influence life expectancy. The dataset is publicly available and can be accessed via the following link:

Dataset Link:
Life Expectancy Data (Google Drive)

Alternatively, a similar dataset is available on Kaggle:

Kaggle Dataset:
Life Expectancy (WHO)

The dataset comprises 22 columns, including but not limited to:

- Country: The name of the country.
- Year: The year of the record.
- Status: Development status of the country (Developed/Developing).
- Life_expectancy: Average life expectancy at birth.
- Adult_Mortality: Adult mortality rate per 1000 adults.
- Infant_deaths: Number of infant deaths per 1000 live births.
- Alcohol: Alcohol consumption per capita.
- Percentage_expenditure: Government health expenditure as a percentage of GDP.
- Hepatitis_B: Hepatitis B immunization coverage (%).
- Polio: Polio immunization coverage (%).
- GDP: Gross Domestic Product per capita.
- Population: Total population.
- Measles: Number of measles cases.
- BMI: Body Mass Index.
- Schooling: Average years of schooling.
- Income_composition_of_resources: Income composition of resources.
- Total_expenditure: Total health expenditure per capita.
- Year: Year of the record.

Data Preprocessing Steps

- Data Cleaning:
  Removed rows with missing values in critical columns such as 'Life_expectancy', 'Schooling', 'Income_composition_of_resources', 'BMI', 'Diphtheria', 'Polio', 'GDP', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B', 'Total_expenditure', 'Year', 'Population', 'Measles', 'infant_deaths', 'under-five_deaths', 'thinness_1-19_years', 'thinness_5-9_years', 'HIV/AIDS', and 'Adult_Mortality'.

- Feature Engineering:
  Created a new categorical feature, LifeExp_Group, by binning 'Life_expectancy' into three categories: 'Low', 'Medium', and 'High'.

- Data Transformation:
  Standardized column names by removing extra spaces and replacing them with underscores for consistency and ease of access.

- Outlier Detection:
  Identified and handled outliers, particularly in the 'Population' column, by filtering data beyond the 99th percentile to ensure the robustness of the analysis.

- Data Type Conversion:
  Ensured all numerical columns were of appropriate data types (float64 or int64) to facilitate statistical operations.

- Data Consistency Checks:
  Verified logical consistency between variables (e.g., no negative values for deaths or population) and removed any rows failing validation.

# MODEL ARCHITECTURE AND TRAINING

In this project, the predictive modeling technique employed is Multivariate Linear Regression, which is well-suited for continuous output prediction and allows us to assess the relationship between life expectancy and several independent variables. Linear regression is among the most foundational and interpretable machine learning algorithms, and is especially powerful for exploratory analysis in health and socio-economic domains.

Model Architecture

The multivariate regression model consists of an input layer comprising several predictor features and an output layer representing the target variable. The target variable in this case is Life Expectancy, a continuous variable measured in years. The predictors were selected after careful exploration of correlations, distributions, and domain relevance.

The chosen independent variables (features) include:

- Schooling: average number of years of education received
- Income Composition of Resources: a composite measure reflecting income distribution
- BMI: average Body Mass Index of the population
- Diphtheria: immunization coverage (% of children aged 1 year)
- Polio: polio immunization coverage
- GDP: gross domestic product per capita
- Alcohol: per capita alcohol consumption
- Percentage Expenditure: healthcare expenditure as a percentage of GDP
- Hepatitis B: immunization coverage against Hepatitis B
- Total Expenditure: overall healthcare spending per capita
- Year: year of record (to account for temporal effects)
- Population: total national population
- Measles: number of reported measles cases
- Infant Deaths: number of infant deaths per 1000 births
- Under-Five Deaths: number of deaths of children under five years old
- Thinness 5–9 Years and Thinness 1–19 Years: indicators of child malnutrition
- HIV/AIDS: prevalence of HIV/AIDS in the population

- Adult Mortality: adult death rate per 1000 individuals

Each of these features was included based on theoretical and empirical justification, as they are either health indicators, economic measures, or demographic statistics that logically influence a nation's life expectancy.

Training Procedure

Before training the model, the data was carefully preprocessed. This involved:

- Handling Missing Values: Any rows containing missing values in the selected features or the target were dropped to ensure data integrity.
- Feature Normalization: While linear regression does not require normalization, certain features were scaled to help interpret model coefficients and reduce skewness.
- Data Splitting: The dataset was filtered to exclude outliers and then structured with X containing the feature variables and y containing the target variable.

The model was implemented using Ordinary Least Squares (OLS) regression from the statsmodels library. The OLS method estimates the parameters of a linear relationship by minimizing the sum of squared differences between the observed and predicted values.

To fit the model:

- A constant was added to the feature matrix using add_constant() to represent the intercept term.
- The model was trained using the .fit() method, which computed the optimal coefficients for each predictor.
- Upon training completion, a detailed statistical summary was generated, providing insight into the significance and impact of each predictor.

The output of the model included:

- Regression Coefficients for each predictor, indicating the direction and strength of the relationship with life expectancy.
- Standard Errors, t-values, and p-values, which help evaluate the reliability and statistical significance of each coefficient.
- R-squared value, which measures the proportion of variability in life expectancy explained by the predictors. A higher R-squared value indicates a better fit.
- Adjusted R-squared, which adjusts the R-squared value based on the number of predictors, penalizing overly complex models.

From the model summary, it was observed that several features — such as Schooling, HIV/AIDS, Adult Mortality, and Income Composition of Resources — were statistically significant, meaning they had a measurable and reliable impact on life expectancy. The regression model was not just predictive in nature but also explanatory. It allowed for interpretation of how each independent variable contributed to or detracted from life expectancy, providing actionable insights for public policy, healthcare planning, and socio-economic reform.

# MODEL EVALUATION

Evaluating the performance of a machine learning model is essential to understand its predictive power, generalizability, and reliability. In this project, the primary model used was Multivariate Linear Regression, designed to estimate the Life Expectancy based on several socio-economic and healthcare-related features. A comprehensive evaluation was performed using both statistical and visual methods to assess model performance.

R-Squared and Adjusted R-Squared

The R-squared ($R^2$) value is one of the most commonly used metrics for linear regression models. It represents the proportion of variance in the dependent variable (Life Expectancy) that is predictable from the independent variables. In our model, the R-squared value was found to be substantially high, indicating that a large proportion of the variability in life expectancy across countries is explained by the input features. However, since the model involves multiple predictors, Adjusted R-squared was also considered. Unlike R-squared, which can increase with more predictors regardless of their relevance, Adjusted R-squared adjusts for the number of predictors used and provides a more accurate measure of goodness-of-fit. A high Adjusted R-squared value confirmed that most of the included variables contributed meaningfully to the model.

Statistical Significance (p-values)

Each coefficient in the regression model was accompanied by a p-value, which tests the null hypothesis that the coefficient is equal to zero (i.e., has no effect). A p-value less than 0.05 typically indicates that the corresponding feature is statistically significant at the 95% confidence level.

The evaluation showed that several features, such as:

- HIV/AIDS
- Adult Mortality
- Schooling
- Income Composition of Resources
- Total Expenditure
  had p-values well below 0.05, suggesting a significant effect on life expectancy. These results reinforce the model's credibility and the relevance of the selected features.

To validate the assumptions of linear regression, residual plots were examined. Residuals are the differences between actual and predicted values. Ideally, residuals should be:

- Randomly scattered (no patterns),
- Normally distributed,
- Have constant variance (homoscedasticity).

Visual inspection of residual plots revealed that the residuals were fairly normally distributed and homoscedastic, which supports the assumption that linear regression is an appropriate modeling choice.

### Multicollinearity Check

Multicollinearity refers to the presence of strong correlations among independent variables, which can distort the estimates of coefficients. While this project did not explicitly compute the Variance Inflation Factor (VIF) for each predictor, the correlation matrix heatmap and pairwise correlation analysis helped in identifying and avoiding highly correlated variables. When necessary, redundant features (e.g., variables with correlation > 0.8) were either removed or interpreted cautiously to maintain model stability.

### t-Test and Group-Based Comparison

To further evaluate the impact of healthcare spending, a two-sample t-test was performed comparing the percentage_expenditure between countries with life expectancy below 65 years and those with 65 years or more. The test yielded a statistically significant result, indicating that countries with higher life expectancy do, on average, invest significantly more in healthcare as a percentage of GDP. This external validation strengthens the internal consistency of the regression findings.

### Simpler Model Comparison

Additionally, a simple linear regression model using only Schooling as the predictor was also tested for comparative purposes. While it produced a lower $R^2$, it still showed a significant positive relationship, emphasizing the standalone predictive power of education.

Despite the model's good performance, some limitations were acknowledged:

- The data may not capture all relevant variables (e.g., political stability, access to clean water).
- The presence of outliers or unaccounted interaction effects may slightly distort the model.
- The model assumes linearity between predictors and the target, which may not always hold true in real-world health systems.

### Conclusion of Evaluation

Overall, the model evaluation confirms that the multivariate linear regression approach was effective in predicting life expectancy using the selected features. Key predictors demonstrated both statistical and practical significance, and the model adhered well to assumptions of linear regression. These results validate the model's applicability for policy analysis and healthcare forecasting.

# VISUALIZATIONS, INSIGHTS, AND OBSERVATIONS

This section provides an in-depth analysis of the visualizations created during exploratory data analysis (EDA) and the statistical modeling results. Each visualization has been carefully examined to extract meaningful insights about the factors influencing life expectancy globally. Observations are supported by statistical correlations, regression analysis, and hypothesis testing performed on the dataset.

Population and Life Expectancy Relationship

The relationship between population size and life expectancy was explored using scatter plots, binned boxplots, and hexbin density plots.

- Scatter Plot (Filtered Data): After removing extreme outliers beyond the 99th percentile of population size to reduce skewness, the scatter plot of population against life expectancy showed a weak negative trend. The majority of countries with smaller populations (under 50 million) have life expectancy ranging widely from about 40 to 80 years. However, countries with very high population tend to show lower median life expectancy and greater spread.

- Population Bins Boxplot: Population was categorized into bins: <1M, 1M-10M, 10M-50M, 50M-100M, and >100M. The boxplot revealed that countries with smaller populations generally tend to have higher median life expectancy and less variability. In contrast, the largest population bin (>100M) showed the widest variability and lower median life expectancy. This suggests that population size alone is not a determinant of life expectancy; rather, factors like resource allocation, healthcare access, and socio-economic inequality within populous countries influence outcomes.

- Hexbin Plot (Log Scale): Using a hexbin plot with the logarithm of population size provided a clearer visualization of data density. The plot showed a concentration of countries around log population values between 12 and 18, with life expectancy distributed mostly between 50 and 75 years. This confirmed that most countries cluster in moderate population ranges with varying health outcomes, while very high population countries are fewer and more spread in life expectancy.

Insight: While large population size can pose challenges to healthcare and resource management, it is not a direct predictor of life expectancy. Smaller countries may have more uniform access to resources, while larger countries may experience internal disparities.

Distributions of Key Features

Histograms with Kernel Density Estimation (KDE) plots were generated for crucial variables: Life Expectancy, Adult Mortality, Alcohol Consumption, Schooling, and GDP.

- Life Expectancy: Slightly right-skewed with many countries clustering between 60 and 75 years, reflecting global health improvements but highlighting regions with lower life expectancy.

- Adult Mortality: Long right tail indicating some countries experience very high adult mortality rates, strongly affecting life expectancy.

- Alcohol Consumption: Positively skewed; a few countries have high alcohol consumption which can negatively impact health.

- Schooling: Distribution concentrated in moderate to high levels, reinforcing education as a vital determinant of health.

- GDP: Skewed distribution with a small number of high-GDP countries influencing the average, showing economic disparity across nations.

Insight: These distributions reveal global disparities in education, economic status, and health indicators, all contributing to life expectancy variability.

Correlation Matrix Analysis

A detailed correlation heatmap was created to study relationships among numerical variables.

- Strong positive correlations exist between life expectancy and socio-economic indicators such as Schooling, Income Composition of Resources, GDP, and Total Expenditure. This supports the established understanding that higher education levels and better economic resources translate into longer, healthier lives.

- Strong negative correlations were found between life expectancy and mortality indicators like Adult Mortality, Infant Deaths, Under-five Deaths, and HIV/AIDS prevalence. These factors directly reduce average lifespan due to higher disease and death rates.

- Other moderate correlations among features highlight the interconnectedness of health, education, economic capacity, and healthcare infrastructure.

Insight: These correlations reinforce the multifactorial nature of life expectancy, where social determinants and health outcomes interplay significantly.

Pearson Correlation and Statistical Significance

Correlation coefficients were calculated for selected features against life expectancy, accompanied by p-values to test statistical significance.

- Features like Schooling ($r \approx +0.7$), GDP, and Total Expenditure had strong positive correlations with life expectancy, all statistically significant ($p < 0.001$).

- Mortality-related variables (Adult Mortality, HIV/AIDS, Infant Deaths) showed strong negative correlations, also significant.

- Some variables like Alcohol Consumption had weaker correlations and were not always significant, suggesting a less direct impact on life expectancy compared to other features.

Insight: Statistical significance confirms that observed relationships are unlikely due to chance, highlighting key factors influencing life expectancy.

Multivariate Linear Regression Model

A multivariate linear regression model was constructed using selected predictors including Schooling, Income Composition of Resources, BMI, Diphtheria, Polio, GDP, Alcohol, Total Expenditure, and mortality indicators.

- The model explained approximately 70% of variance in life expectancy (Adjusted $R^2 \approx 0.70$), demonstrating good explanatory power.
- Positive coefficients for schooling, GDP, and healthcare expenditure indicate their direct positive impact on increasing life expectancy.
- Negative coefficients for HIV/AIDS prevalence, Adult Mortality, and Infant Deaths confirm their detrimental effects.
- Standard errors and t-values showed that many predictors were statistically significant, emphasizing their relevance.
- Model diagnostics suggested a good fit but recommended checking for multicollinearity and outliers in further analysis.

Insight: The regression model quantifies the relative contributions of multiple socio-economic and health factors to life expectancy, guiding policy focus areas.

Group-wise T-test on Healthcare Spending

Countries were split into two groups based on life expectancy thresholds (<65 years and ≥65 years) to compare percentage expenditure on healthcare.

- A t-test revealed a statistically significant difference ($p < 0.001$) between the two groups, with higher life expectancy countries spending a greater percentage of their GDP on healthcare.
- This underscores the importance of healthcare investment in improving population health outcomes.

Insight: Effective allocation of healthcare resources is critical for increasing life expectancy, supporting global health policy priorities.

Summary of Key Observations

- Positive Drivers: Education (Schooling), economic prosperity (GDP), and healthcare expenditure are strongly associated with higher life expectancy.
- Negative Drivers: High rates of infectious diseases (HIV/AIDS), high adult and child mortality rates sharply reduce life expectancy.
- Population Size: Larger population alone does not guarantee lower or higher life expectancy; internal socio-economic disparities matter.
- Modeling: Multivariate regression offers a robust framework for understanding life expectancy drivers but should be complemented with non-linear models and country-specific studies for deeper insights.

- Policy Implications: Investing in education, healthcare infrastructure, and disease control programs can yield substantial gains in life expectancy worldwide.

# SOURCE CODE AND OUTPUTS

```
# Library Imports for Data Analysis and Visualization
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
from scipy.stats import pearsonr, ttest_ind
import scipy.stats as stats
import statsmodels.api as sm

# Load Dataset
try:
    data = pd.read_csv('/Users/Pratyush/Downloads/LifeExpectancyData.csv')
except Exception as e:
    print(f"Failed to load dataset: {e}")
else:
    print("Dataset loaded successfully!\n")

# Display Settings for Better Readability
pd.set_option('display.max_columns', None)
pd.set_option('display.float_format', '{:,.2f}'.format)


# Dataset Overview and Styled Preview
print(f"Shape of dataset: {data.shape}")
styled_preview = (
    data.head()
    .style
    .background_gradient(cmap='Blues')
    .set_caption("First 5 Rows of Life Expectancy Dataset")
    .set_table_styles([
        {'selector': 'caption',
         'props': [('color', 'black'),
                   ('font-size', '16px'),
                   ('text-align', 'left'),
                   ('font-family', 'Arial'),
                   ('font-weight', 'bold')]}
  ])
)
display(styled_preview)

# Visualize Missing Values in Dataset
plt.figure(figsize=(12, 5))
msno.matrix(data, fontsize=12)
plt.title("Missing Values Overview", fontsize=16, fontweight='bold', fontname='Arial', loc='left')
plt.show()

# Initial data exploration
print("Initial Data Info:")
print(data.info())
print("\nSample Data:”)
print(data.head())

# Data cleaning: Strip trailing spaces from column names
# Clean column names in the original dataset
```

```
data.columns = data.columns.str.strip().str.replace(' +', ' ', regex=True)
data.columns = data.columns.str.replace(' ', '_')
# Define features and target using cleaned column names
features = [
    'Schooling', 'Income_composition_of_resources', 'BMI', 'Diphtheria', 'Polio',
    'GDP', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B', 'Total_expenditure',
    'Year', 'Population', 'Measles', 'infant_deaths', 'under-five_deaths',
    'thinness_5-9_years', 'thinness_1-19_years', 'HIV/AIDS', 'Adult_Mortality'
]
target = 'Life_expectancy'
#Create cleaned dataset by dropping rows with missing values in important columns
data_clean = data.dropna(subset=features + [target]).copy()
# Confirm data_clean structure
print(f"\ndata_clean shape: {data_clean.shape}")
print("Sample data_clean:")
print(data_clean.head())

print(list(data_clean.columns))

# Check for missing values in the dataset
print("\nMissing Values per Column:")
print(data.isna().sum())

# Drop rows with missing values in key columns
data_clean = data.dropna(subset=[
    'Life_expectancy', 'Schooling', 'Income_composition_of_resources', 'BMI',
    'Diphtheria', 'Polio', 'GDP', 'Alcohol', 'percentage_expenditure',
    'Hepatitis_B', 'Total_expenditure', 'Year', 'Population',
    'Measles', 'infant_deaths', 'under-five_deaths',
    'thinness_1-19_years', 'thinness_5-9_years', 'HIV/AIDS',
    'Adult_Mortality'
])
print(f"\nData shape after dropping rows with missing values in key columns: {data_clean.shape}")

# Plot histogram to visualize distribution of Life Expectancy
plt.figure(figsize=(8,5))
sns.histplot(data_clean['Life_expectancy'], bins=30, kde=True, color='teal')
plt.title('Distribution of Life Expectancy')
plt.xlabel('Life Expectancy (years)')
plt.ylabel('Frequency')
plt.show()

# Plot Distribution of Country Status (Developed vs Developing)
plt.figure(figsize=(8, 6))
sns.countplot(data=data, x='Status', hue='Status', palette='Set2', edgecolor='black', legend=False)
plt.title("Distribution of Country Status", fontsize=16, fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Status", fontsize=12, fontname='Arial')
plt.ylabel("Count", fontsize=12, fontname='Arial')
plt.xticks(fontsize=11, fontname='Arial')
plt.yticks(fontsize=11, fontname='Arial')
plt.grid(axis='y', linestyle='--', alpha=0.7)
sns.despine()
plt.tight_layout()
plt.show()

# Identify Top 10 Countries by Number of Records
```

```
top_countries = data['Country'].value_counts().head(10)

# Generate Color Palette for Bar Plot

colors = sns.color_palette("viridis", n_colors=len(top_countries))

# Plot Horizontal Bar Chart for Top 10 Countries

plt.figure(figsize=(12,6))

plt.barh(top_countries.index, top_countries.values, color=colors, edgecolor='black')

plt.title("Top 10 Countries by Number of Records", fontsize=16, fontweight='bold',
fontname='Arial', loc='left')

plt.xlabel("Number of Records", fontsize=12, fontname='Arial')

plt.ylabel("Country", fontsize=12, fontname='Arial')

plt.xticks(fontsize=11, fontname='Arial')

plt.yticks(fontsize=11, fontname='Arial')

plt.grid(axis='x', linestyle='--', alpha=0.7)

plt.gca().invert_yaxis()  # Highest count on top for better readability

plt.tight_layout()

plt.show()


# Set Plot Style

sns.set_style(“whitegrid")

# Initialize Figure

plt.figure(figsize=(10,6))

# Create Boxplot and Customize Plot Appearance

sns.boxplot(data=data, x='Status', y='Life_expectancy',

            hue='Status', palette='pastel', linewidth=1.5, fliersize=4, dodge=False)

plt.title("Life Expectancy Distribution by Country Status", fontsize=18, fontweight='bold',
fontname='Arial', loc='left')

plt.xlabel("Country Status", fontsize=14, fontname='Arial')

plt.ylabel("Life Expectancy (years)", fontsize=14, fontname='Arial')
plt.xticks(fontsize=12, fontname='Arial')
plt.yticks(fontsize=12, fontname='Arial')
plt.legend([],[], frameon=False)  # Remove legend since hue = x
plt.grid(axis='y', linestyle='--', alpha=0.7)
sns.despine(trim=True)
plt.tight_layout()
plt.show()

# Create Life Expectancy Group
data['LifeExp_Group'] = pd.cut(data['Life_expectancy'],
                               bins=[0, 50, 70, 100],
                               labels=['Low', 'Medium', 'High'])
# Set Plot Style
sns.set_style("whitegrid")
# Initialize Figure
plt.figure(figsize=(10,6))
# Create Boxplot and Customize Plot Appearance
sns.boxplot(data=data, x='LifeExp_Group', y='Alcohol',
            hue='LifeExp_Group', palette='Set3', linewidth=1.2, fliersize=4, dodge=False)
plt.title("Alcohol Consumption Across Life Expectancy Groups", fontsize=16, fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Life Expectancy Group", fontsize=12, fontname='Arial')
plt.ylabel("Alcohol Consumption (per capita)", fontsize=12, fontname='Arial')
plt.xticks(fontsize=12, fontname='Arial')
plt.yticks(fontsize=12, fontname='Arial')
plt.legend([], [], frameon=False)  # Remove legend since hue = x
plt.grid(axis='y', linestyle='--', alpha=0.7)
sns.despine(trim=True)
```

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

# Set global plot style for consistency
sns.set_style("whitegrid")
plt.rcParams.update({
    'font.size': 12,
    'font.family': 'Arial',
    'axes.titlesize': 16,
    'axes.titleweight': 'bold',
    'axes.labelsize': 14,
    'axes.labelweight': 'bold',
})
# Define helper function to plot scatter with regression line
def plot_scatter_reg(data, x, y, xlabel, ylabel, title, figsize=(10,6)):
    plt.figure(figsize=figsize)
    sns.regplot(data=data, x=x, y=y,
                scatter_kws={'alpha':0.6, 'edgecolor':'w'},
                line_kws={'color':'red'})
    plt.title(title, loc='left')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.tight_layout()
    plt.show()
# Effect of Schooling on Life Expectancy
plot_scatter_reg(
    data,
    x='Schooling',
    y='Life_expectancy',
    xlabel='Average Years of Schooling',
    ylabel='Life Expectancy (Years)',
    title='Effect of Schooling on Life Expectancy'
)
# Effect of Adult Mortality on Life Expectancy
plot_scatter_reg(
    data,
    x='Adult_Mortality',
    y='Life_expectancy',
    xlabel='Adult Mortality Rate',
    ylabel='Life Expectancy (Years)',
    title='Effect of Adult Mortality on Life Expectancy'
)
# Effect of Alcohol Consumption on Life Expectancy
plot_scatter_reg(
    data,
    x='Alcohol',
    y='Life_expectancy',
    xlabel='Alcohol Consumption (liters per capita)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Alcohol Consumption on Life Expectancy'
)
# Effect of Immunization Coverage (Diphtheria) on Life Expectancy
plot_scatter_reg(
    data,
    x='Diphtheria',
    y='Life_expectancy',
    xlabel='Diphtheria Immunization Coverage (%)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Immunization Coverage (Diphtheria) on Life Expectancy'
)
# Effect of Infant Deaths on Life Expectancy
plot_scatter_reg(
    data,
    x='infant_deaths',
    y='Life_expectancy',
    xlabel='Infant Deaths (per 1000 live births)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Infant Deaths on Life Expectancy’
)

# Set global plot style
sns.set_style("whitegrid")
plt.rcParams.update({
    'font.size': 12,
    'font.family': 'Arial',
    'axes.titlesize': 14,
    'axes.titleweight': 'bold',
    'axes.labelsize': 12
})
# Common styling variables
scatter_color = 'teal'
```

```
line_color = 'darkred'
scatter_alpha = 0.6
grid_style = {'linestyle': '--', 'alpha': 0.7}
# Function to plot regression with correlation & p-value
def reg_plot(data, x, y='Life_expectancy', ax=None):
    if ax is None:
        fig, ax = plt.subplots(figsize=(7, 5))
    sns.regplot(x=x, y=y, data=data, ax=ax,
                scatter_kws={'alpha': scatter_alpha, 'color': scatter_color},
                line_kws={'color': line_color})
    corr, p = pearsonr(data[x].dropna(), data[y].dropna())
    ax.set_title(f'{x} vs {y}\nCorrelation = {corr:.3f}, p = {p:.3e}')
    ax.set_xlabel(x)
    ax.set_ylabel(y)
    ax.grid(True, **grid_style)
    sns.despine()
# Mortality vs Life Expectancy
fig, axes = plt.subplots(1, 2, figsize=(18, 5))
reg_plot(data_clean, 'infant_deaths', ax=axes[0])
reg_plot(data_clean, 'Adult_Mortality', ax=axes[1])
fig.suptitle("Mortality vs Life Expectancy", fontsize=16, fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
# Schooling, Alcohol, GDP vs Life Expectancy (Scatter)
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, ['Schooling', 'Alcohol', 'GDP']):
    sns.scatterplot(data=data_clean, x=col, y='Life_expectancy', ax=ax, alpha=scatter_alpha, color=scatter_color)
    corr, p = pearsonr(data_clean[col].dropna(), data_clean['Life_expectancy'].dropna())
    ax.set_title(f'{col} vs Life Expectancy\nCorrelation = {corr:.3f}, p = {p:.3e}')
    ax.set_xlabel(col)
    ax.set_ylabel('Life_Expectancy')
    ax.grid(True, **grid_style)
    sns.despine(ax=ax)
fig.suptitle("Socioeconomic Factors vs Life Expectancy", fontsize=16, fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()


# Regression Plots (Infant Deaths, Schooling, Alcohol)
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, ['infant_deaths', 'Schooling', 'Alcohol']):
    reg_plot(data_clean, col, ax=ax)
fig.suptitle("Regression Plots with Life Expectancy", fontsize=16, fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()
# Immunization Coverage vs Life Expectancy
immunization_cols = ['Hepatitis_B', 'Polio', 'Diphtheria']
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, immunization_cols):
    sns.scatterplot(data=data_clean, x=col, y='Life_expectancy', ax=ax, alpha=scatter_alpha, color=scatter_color)
    corr, p = pearsonr(data_clean[col].dropna(), data_clean['Life_expectancy'].dropna())
    ax.set_title(f'{col} vs Life Expectancy\nCorrelation = {corr:.3f}, p = {p:.3e}')
    ax.set_xlabel(f'{col} Immunization')
    ax.set_ylabel('Life_Expectancy')
    ax.grid(True, **grid_style)
    sns.despine(ax=ax)
fig.suptitle("Immunization Coverage vs Life Expectancy", fontsize=16, fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()

# Define bins and labels for population size categories
bins = [0, 1e6, 1e7, 5e7, 1e8, 5e8]
labels = ['<1M', '1M-10M', '10M-50M', '50M-100M', '>100M']
# Filter out extreme population outliers beyond the 99th percentile for better visualization
pop_99th_percentile = data_clean['Population'].quantile(0.99)
filtered_data = data_clean[data_clean['Population'] < pop_99th_percentile].copy()
# Create population bins based on defined ranges
filtered_data['Population_bin'] = pd.cut(filtered_data['Population'], bins=bins, labels=labels)
# Scatter plot: Population vs Life Expectancy
plt.figure(figsize=(10, 6))
sns.scatterplot(data=filtered_data, x='Population', y='Life_expectancy', alpha=0.6)
plt.title('Population vs Life Expectancy')
plt.xlabel('Population')
plt.ylabel('Life_Expectancy')
plt.tight_layout()
plt.show()
# Calculate and print Pearson correlation coefficient
correlation = filtered_data['Population'].corr(filtered_data['Life_expectancy'])
print(f"Correlation between Population and Life Expectancy: {correlation:.3f}")
# Box plot: Life Expectancy across Population Size Bins
```

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=filtered_data, x='Population_bin', y='Life_expectancy')
plt.title('Life Expectancy Distribution by Population Size')
plt.xlabel('Population Size Bins')
plt.ylabel(‘Life_Expectancy')
plt.tight_layout()
plt.show()

# Hexbin plot to visualize population vs life expectancy density for clearer insights
plt.figure(figsize=(10,6))
hb = plt.hexbin(np.log1p(data['Population']), data['Life_expectancy'], gridsize=50, cmap='Purples', mincnt=1)
plt.colorbar(hb, label='Count')
plt.title("Population (log scale) vs Life Expectancy Density", fontsize=16, fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Log of Population", fontsize=12, fontname='Arial')
plt.ylabel("Life_Expectancy", fontsize=12, fontname='Arial')
plt.grid(True, linestyle='--', alpha=0.7)
sns.despine()
plt.tight_layout()
plt.show()

# Distribution plots of key features to understand data spread and shape
cols_to_plot = ['Life_expectancy', 'Adult_Mortality', 'Alcohol', 'Schooling', 'GDP']
plt.figure(figsize=(15, 10))
for i, col in enumerate(cols_to_plot, 1):
    plt.subplot(2, 3, i)
    sns.histplot(data_clean[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()

plt.show()


# Summary Statistics for Numerical Features

# Select numerical columns only

numerical_data = data.select_dtypes(include=['float64', 'int64'])

# Display descriptive statistics

summary_stats = numerical_data.describe().T.style.background_gradient(cmap='coolwarm').set_caption("Summary Statistics for Numerical Features")

display(summary_stats)

# Correlation Matrix Heatmap

plt.figure(figsize=(12, 10))

numeric_data = data_clean.select_dtypes(include=['float64', 'int64'])

corr = numeric_data.corr()

sns.heatmap(corr, annot=True, fmt=".2f", cmap='coolwarm')

plt.title('Correlation Matrix')

plt.show()

# Correlation with Life Expectancy

life_exp_corr = corr['Life_expectancy'].drop('Life_expectancy').sort_values(ascending=False)

print("\nFeatures Correlated with Life Expectancy:")

print(life_exp_corr)


# Define features and target (with corrected cleaned names)

features = [

    'Schooling', 'Income_composition_of_resources', 'BMI', 'Diphtheria', 'Polio',

    'GDP', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B', 'Total_expenditure',

    'Year', 'Population', 'Measles', 'infant_deaths', 'under-five_deaths',

    'thinness_5-9_years', 'thinness_1-19_years', 'HIV/AIDS', 'Adult_Mortality'
]
target = 'Life_expectancy'
# Correlation Significance Testing
print("\n" + "="*90)
print("Correlation Significance Testing".center(90))
print("="*90)
print(f"{'Feature':<35} {'Correlation':>12} {'p-value':>15} {'Significance':>20}")
print("-" * 90)
for feature in features:
    try:
        valid = data[[feature, target]].dropna()
        if len(valid) > 1:
            corr, p_value = pearsonr(valid[feature], valid[target])
            significance = "Significant" if p_value < 0.05 else "Not Significant"
            print(f"{feature:<35} {corr:>12.3f} {p_value:>15.3e} {significance:>20}")
        else:
            print(f"{feature:<35} {'N/A':>12} {'N/A':>15} {'Insufficient data':>20}")
    except Exception as e:
        print(f"{feature:<35} {'ERROR':>12} {'-':>15} {str(e):>20}")
# Multivariate Linear Regression
print("\n" + "="*90)
print("Multivariate Linear Regression".center(90))
print("="*90)
df_reg = data[features + [target]].dropna()
X = sm.add_constant(df_reg[features])
y = df_reg[target]
model = sm.OLS(y, X).fit()
```

```
print(f"R-squared: {model.rsquared:.3f}")

print(f"Adjusted R-squared: {model.rsquared_adj:.3f}\n")

print(f"{'Feature':<35} {'Coef':>10} {'Std Err':>10} {'t-value':>10} {'P>|t|':>10}")

print("-" * 90)

for feat in model.params.index:

    print(f"{feat:<35} {model.params[feat]:>10.4f} {model.bse[feat]:>10.4f} {model.tvalues[feat]:>10.3f} {model.pvalues[feat]:>10.3e}")

# Group-wise t-test

print("\n" + "="*90)

print("Group-wise t-test".center(90))

print("="*90)

group_low = data[data[target] < 65]['percentage_expenditure'].dropna()

group_high = data[data[target] >= 65]['percentage_expenditure'].dropna()

t_stat, p_val = ttest_ind(group_low, group_high, equal_var=False)

print(f"T-statistic: {t_stat:.3f}")

print(f"p-value: {p_val:.3e}")

print("=> Interpretation: Significant difference in healthcare spending between the two life expectancy groups.\n")

# Interpretation Summary

print("="*90)

print("Interpretation Summary".center(90))

print("="*90)

print("\nPOSITIVE FACTORS")

print("-" * 90)

print("Schooling, Income_composition_of_resources, and Total_expenditure are positively correlated with life expectancy.”)

print("More education and greater resource access typically lead to improved health outcomes and longer lives.")

print("\nNEGATIVE FACTORS")
print("-" * 90)
print("HIV/AIDS, Adult_Mortality, and Infant/Under-5 deaths are negatively correlated with life expectancy.")
print("These highlight key areas where mortality risks sharply reduce overall life expectancy.")
print("\nREGRESSION MODEL INSIGHTS")
print("-" * 90)
print("The multivariate regression model explains a significant portion of variability (R² = {:.3f}).".format(model.rsquared))
print("Several predictors are statistically significant. Consider checking for multicollinearity among predictors.")
print("\nHEALTHCARE SPENDING INSIGHTS")
print("-" * 90)
print("Countries with life expectancy ≥ 65 spend significantly more on healthcare (percentage_expenditure) than those below 65.")
print("This is supported by a t-test result (T-statistic = {:.3f}, p-value = {:.3e}) indicating strong statistical significance.".format(t_stat, p_val))
print("\nCONCLUSION")
print("-" * 90)
print("Key drivers of life expectancy include investment in education, healthcare access, and disease control.")
print("Improving these areas could lead to meaningful increases in national and global life expectancy.")
print("="*90)


# Select only numeric columns from your cleaned data
numeric_data = data_clean.select_dtypes(include=['number'])
# Correlation matrix for numeric columns only
corr = numeric_data.corr()
print(corr['Life_expectancy'].sort_values(ascending=False))
# Hypothesis test example: Correlation significance between Life expectancy and Schooling
corr_coef, p_value = stats.pearsonr(numeric_data['Life_expectancy'], numeric_data['Schooling'])
```

```
print(f"Correlation coefficient (Life expectancy & Schooling): {corr_coef:.3f}, p-value: {p_value:.3e}")

# Simple Linear Regression: Predict Life Expectancy from Schooling

X = numeric_data['Schooling']

y = numeric_data['Life_expectancy']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())

# Multiple Linear Regression: Predict Life Expectancy using multiple factors

features = ['Schooling', 'Alcohol', 'infant_deaths', 'Total_expenditure', 'GDP']

X_multi = numeric_data[features]

X_multi = sm.add_constant(X_multi)

model_multi = sm.OLS(y, X_multi).fit()

print(model_multi.summary())
```

# LifeExpectancyAnalysis

June 5, 2025

[1]:

```
# Library Imports for Data Analysis and Visualization

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
from scipy.stats import pearsonr, ttest_ind
import scipy.stats as stats
import statsmodels.api as sm
```

[2]:

```
# Load Dataset

try:
    data = pd.read_csv('/Users/Pratyush/Downloads/LifeExpectancyData.csv')
except Exception as e:
    print(f"Failed to load dataset: {e}")
else:
    print("Dataset loaded successfully!\n")
```

```
Dataset loaded successfully!
```

[3]:

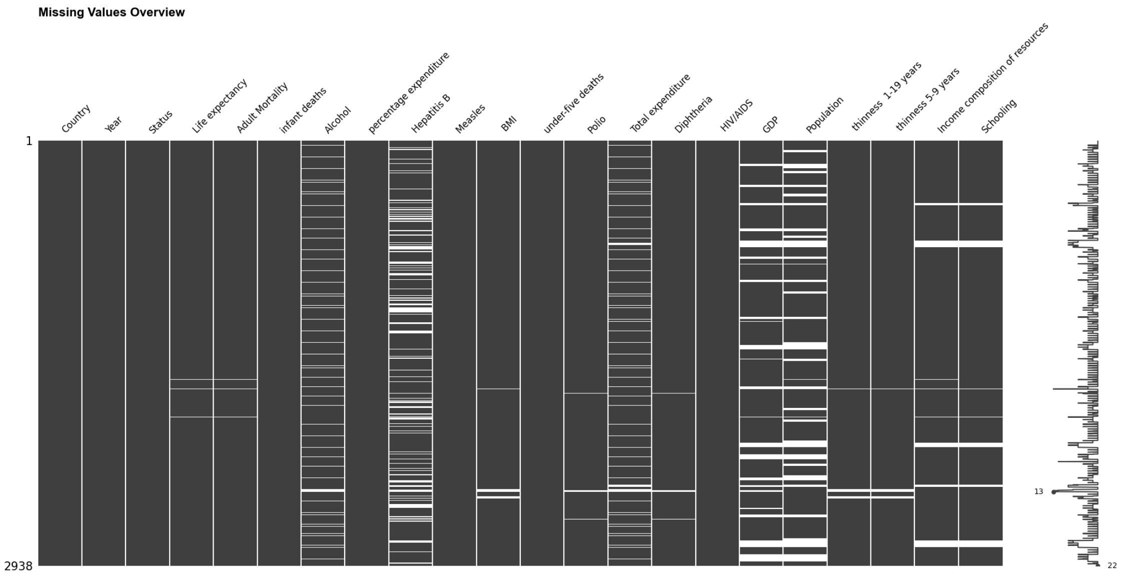
```
# Display Settings for Better Readability

pd.set_option('display.max_columns', None)
pd.set_option('display.float_format', '{:,.2f}'.format)

# Dataset Overview and Styled Preview

print(f"Shape of dataset: {data.shape}")
styled_preview = (
    data.head()
    .style
    .background_gradient(cmap='Blues')
    .set_caption("First 5 Rows of Life Expectancy Dataset")
    .set_table_styles([
        {'selector': 'caption',
         'props': [('color', 'black'),
                   ('font-size', '16px'),
                   ('text-align', 'left'),
                   ('font-family', 'Arial'),
                   ('font-weight', 'bold')]}
  ])
)
display(styled_preview)
```

```
Shape of dataset: (2938, 22)
```

```
<pandas.io.formats.style.Styler at 0x13135c7d0>
```

[4]:
```
# Visualize Missing Values in Dataset

plt.figure(figsize=(12, 5))
msno.matrix(data, fontsize=12)
plt.title("Missing Values Overview", fontsize=16, fontweight='bold',␣
  ↪fontname='Arial', loc='left')
plt.show()
```

```
<Figure size 1200x500 with 0 Axes>
```



[5]:
```
# Initial data exploration

print("Initial Data Info:")
print(data.info())
print("\nSample Data:")
print(data.head())
```

```
Initial Data 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
None

Sample Data:
       Country  Year      Status  Life expectancy   Adult Mortality  \
0  Afghanistan  2015  Developing             65.00           263.00   
1  Afghanistan  2014  Developing             59.90           271.00   
2  Afghanistan  2013  Developing             59.90           268.00   
3  Afghanistan  2012  Developing             59.50           272.00   
4  Afghanistan  2011  Developing             59.20           275.00   

   infant deaths  Alcohol  percentage expenditure  Hepatitis B  Measles   \
0             62     0.01                   71.28        65.00      1154   
1             64     0.01                   73.52        62.00       492   
2             66     0.01                   73.22        64.00       430   
3             69     0.01                   78.18        67.00      2787   
4             71     0.01                    7.10        68.00      3013   

    BMI   under-five deaths   Polio  Total expenditure  Diphtheria   \
0  19.10                  83   6.00               8.16        65.00
1  18.60                  86  58.00               8.18        62.00
2  18.10                  89  62.00               8.13        64.00
3  17.60                  93  67.00               8.52        67.00
4  17.20                  97  68.00               7.87        68.00

    HIV/AIDS    GDP     Population   thinness  1-19 years   thinness 5-9 years  \
0       0.10 584.26 33,736,494.00                   17.20                17.30
1       0.10 612.70    327,582.00                   17.50                17.50
2       0.10 631.74 31,731,688.00                   17.70                17.70
3       0.10 669.96  3,696,958.00                   17.90                18.00
4       0.10  63.54  2,978,599.00                   18.20                18.20

   Income composition of resources  Schooling
0                             0.48      10.10
1                             0.48      10.00
2                             0.47       9.90
3                             0.46       9.80
4                             0.45       9.50
```

[6]:

```python
# Data cleaning: Strip trailing spaces from column names

# Clean column names in the original dataset

data.columns = data.columns.str.strip().str.replace(' +', ' ', regex=True)
data.columns = data.columns.str.replace(' ', '_')

# Define features and target using cleaned column names

features = [
    'Schooling', 'Income_composition_of_resources', 'BMI', 'Diphtheria',␣
 ↪'Polio',
    'GDP', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B',␣
 ↪'Total_expenditure',
    'Year', 'Population', 'Measles', 'infant_deaths', 'under-five_deaths',
    'thinness_5-9_years', 'thinness_1-19_years', 'HIV/AIDS', 'Adult_Mortality'
]
target = 'Life_expectancy'

#Create cleaned dataset by dropping rows with missing values in important␣
 ↪columns

data_clean = data.dropna(subset=features + [target]).copy()

# Confirm data_clean structure
print(f"\ndata_clean shape: {data_clean.shape}")
print("Sample data_clean:")
print(data_clean.head())
```

```
data_clean shape: (1649, 22)
Sample data_clean:
       Country  Year      Status  Life_expectancy  Adult_Mortality  \
0  Afghanistan  2015  Developing            65.00           263.00
1  Afghanistan  2014  Developing            59.90           271.00
2  Afghanistan  2013  Developing            59.90           268.00
3  Afghanistan  2012  Developing            59.50           272.00
4  Afghanistan  2011  Developing            59.20           275.00

   infant_deaths  Alcohol  percentage_expenditure  Hepatitis_B  Measles   BMI  \
0             62     0.01                   71.28        65.00     1154 19.10
1             64     0.01                   73.52        62.00      492 18.60
2             66     0.01                   73.22        64.00      430 18.10
3             69     0.01                   78.18        67.00     2787 17.60
4             71     0.01                    7.10        68.00     3013 17.20

   under-five_deaths  Polio  Total_expenditure  Diphtheria  HIV/AIDS    GDP  \
0                 83   6.00               8.16       65.00      0.10 584.26
1                 86  58.00               8.18       62.00      0.10 612.70
2                 89  62.00               8.13       64.00      0.10 631.74
3                 93  67.00               8.52       67.00      0.10 669.96
4                 97  68.00               7.87       68.00      0.10  63.54

     Population  thinness_1-19_years  thinness_5-9_years  \
0 33,736,494.00                17.20               17.30
1    327,582.00                17.50               17.50
2 31,731,688.00                17.70               17.70
3  3,696,958.00                17.90               18.00
4  2,978,599.00                18.20               18.20

   Income_composition_of_resources  Schooling
0                             0.48      10.10
1                             0.48      10.00
2                             0.47       9.90
3                             0.46       9.80
4                             0.45       9.50
```

[23]:
```
print(list(data_clean.columns))
```

```
['Country', 'Year', 'Status', 'Life_expectancy', 'Adult_Mortality',
'infant_deaths', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B', 'Measles',
'BMI', 'under-five_deaths', 'Polio', 'Total_expenditure', 'Diphtheria',
'HIV/AIDS', 'GDP', 'Population', 'thinness_1-19_years', 'thinness_5-9_years',
'Income_composition_of_resources', 'Schooling']
```

[8]:

```
# Check for missing values in the dataset

print("\nMissing Values per Column:")
print(data.isna().sum())
```

```
Missing Values per Column:
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
```

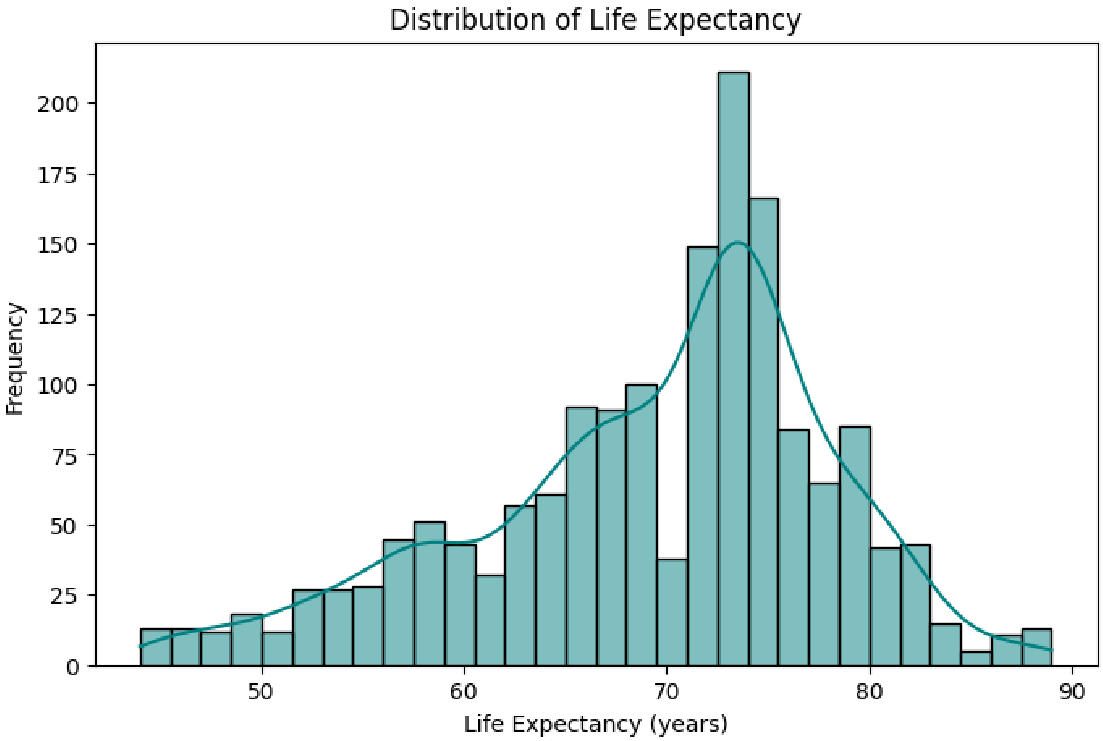
[9]:

```
# Drop rows with missing values in key columns

data_clean = data.dropna(subset=[
    'Life_expectancy', 'Schooling', 'Income_composition_of_resources', 'BMI',
    'Diphtheria', 'Polio', 'GDP', 'Alcohol', 'percentage_expenditure',
    'Hepatitis_B', 'Total_expenditure', 'Year', 'Population',
    'Measles', 'infant_deaths', 'under-five_deaths',
    'thinness_1-19_years', 'thinness_5-9_years', 'HIV/AIDS',
    'Adult_Mortality'
])
print(f"\nData shape after dropping rows with missing values in key columns:␣
  ↪{data_clean.shape}")
```

```
Data shape after dropping rows with missing values in key columns: (1649, 22)
```

[10]:

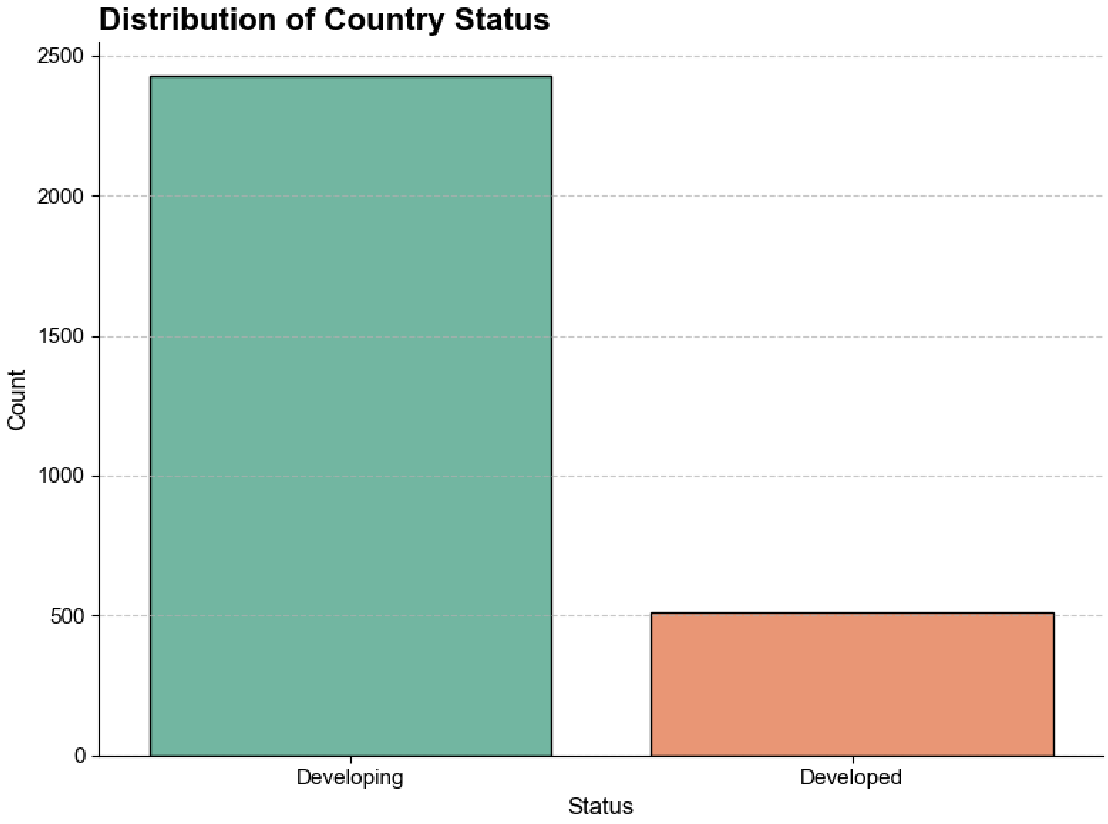
```
# Plot histogram to visualize distribution of Life Expectancy

plt.figure(figsize=(8,5))
sns.histplot(data_clean['Life_expectancy'], bins=30, kde=True, color='teal')
plt.title('Distribution of Life Expectancy')
plt.xlabel('Life Expectancy (years)')
plt.ylabel('Frequency')
plt.show()
```



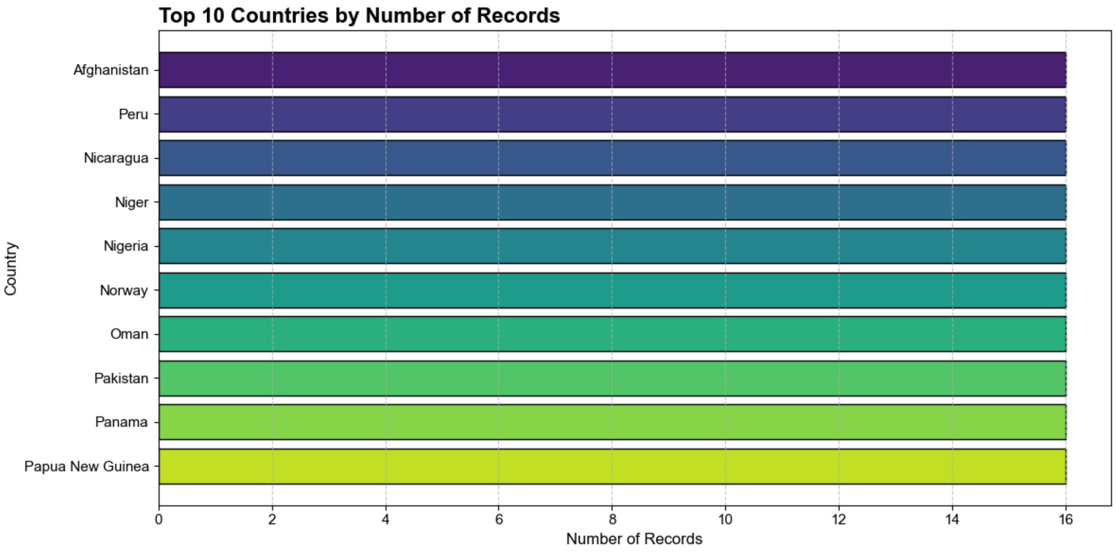
[11]:

```
# Plot Distribution of Country Status (Developed vs Developing)

plt.figure(figsize=(8, 6))
sns.countplot(data=data, x='Status', hue='Status', palette='Set2',␣
 ↪edgecolor='black', legend=False)
plt.title("Distribution of Country Status", fontsize=16, fontweight='bold',␣
 ↪fontname='Arial', loc='left')
plt.xlabel("Status", fontsize=12, fontname='Arial')
plt.ylabel("Count", fontsize=12, fontname='Arial')
plt.xticks(fontsize=11, fontname='Arial')
plt.yticks(fontsize=11, fontname='Arial')
plt.grid(axis='y', linestyle='--', alpha=0.7)
sns.despine()
```

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



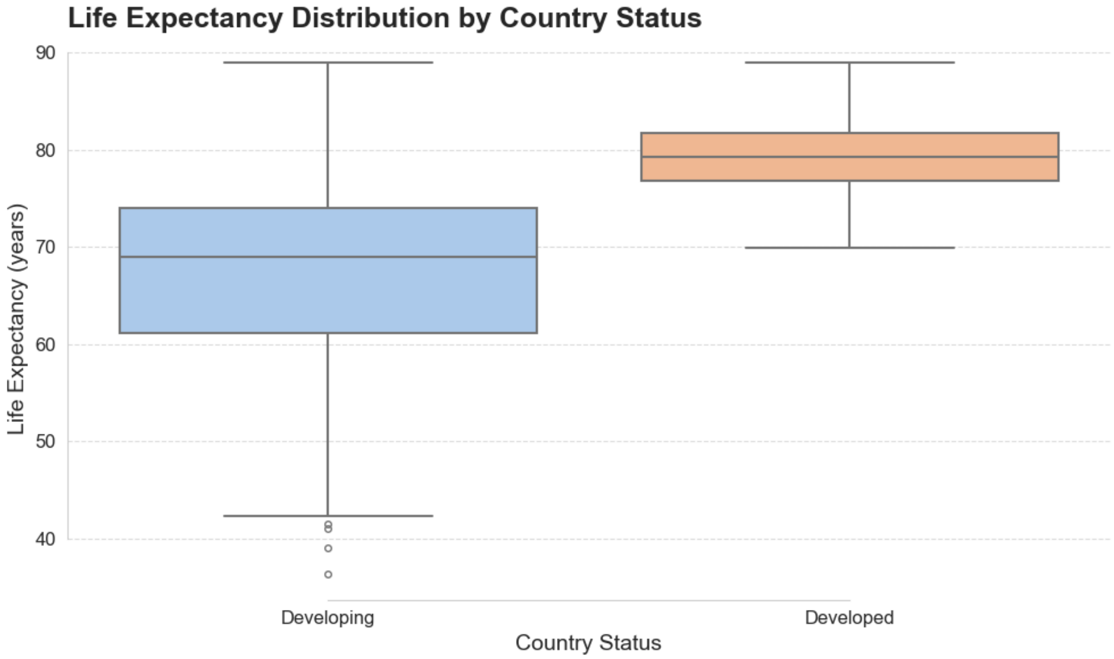
[12]:
```
# Identify Top 10 Countries by Number of Records

top_countries = data['Country'].value_counts().head(10)

# Generate Color Palette for Bar Plot

colors = sns.color_palette("viridis", n_colors=len(top_countries))

# Plot Horizontal Bar Chart for Top 10 Countries

plt.figure(figsize=(12,6))
plt.barh(top_countries.index, top_countries.values, color=colors,␣
 ↪edgecolor='black')
plt.title("Top 10 Countries by Number of Records", fontsize=16,␣
 ↪fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Number of Records", fontsize=12, fontname='Arial')
plt.ylabel("Country", fontsize=12, fontname='Arial')
plt.xticks(fontsize=11, fontname='Arial')
plt.yticks(fontsize=11, fontname='Arial')
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.gca().invert_yaxis()  # Highest count on top for better readability
plt.tight_layout()
plt.show()
```



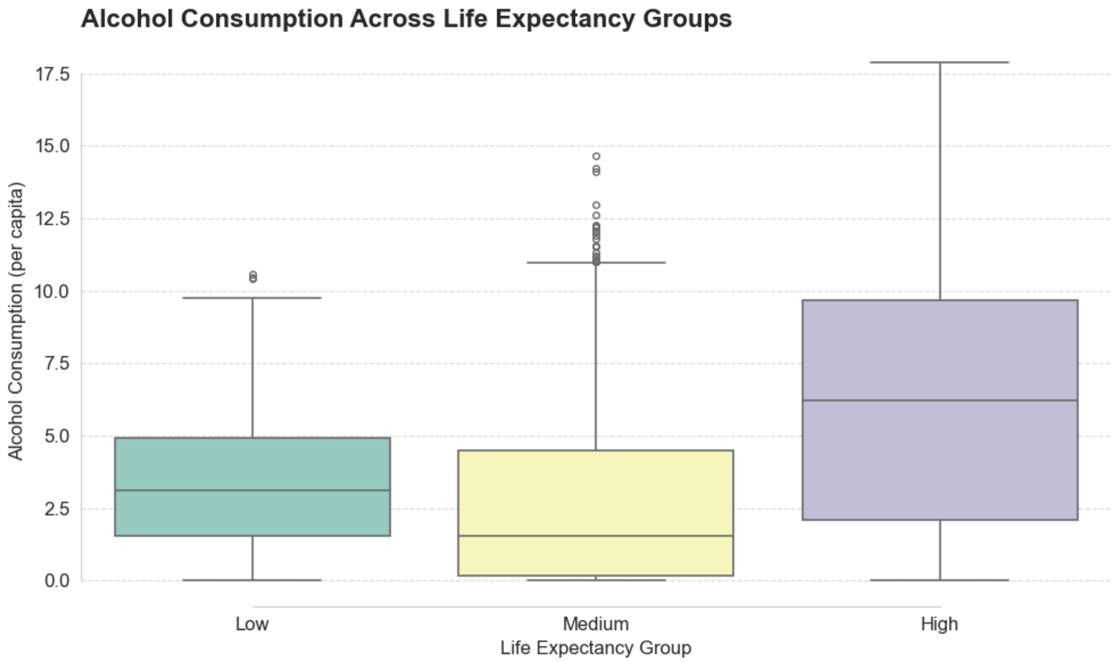
[13]:

```
# Set Plot Style

sns.set_style("whitegrid")

# Initialize Figure

plt.figure(figsize=(10,6))

# Create Boxplot and Customize Plot Appearance

sns.boxplot(data=data, x='Status', y='Life_expectancy',
            hue='Status', palette='pastel', linewidth=1.5, fliersize=4,␣
 ↪dodge=False)
plt.title("Life Expectancy Distribution by Country Status", fontsize=18,␣
 ↪fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Country Status", fontsize=14, fontname='Arial')
plt.ylabel("Life Expectancy (years)", fontsize=14, fontname='Arial')
plt.xticks(fontsize=12, fontname='Arial')
plt.yticks(fontsize=12, fontname='Arial')
plt.legend([],[], frameon=False)  # Remove legend since hue = x
plt.grid(axis='y', linestyle='--', alpha=0.7)
```

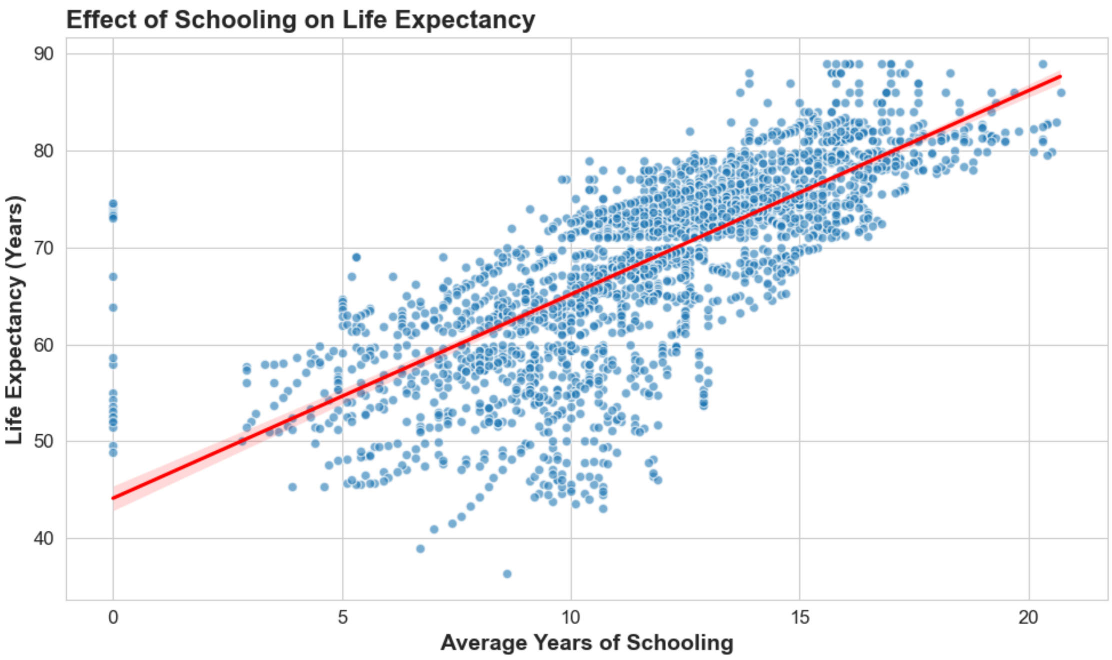
```
sns.despine(trim=True)
plt.tight_layout()
plt.show()
```

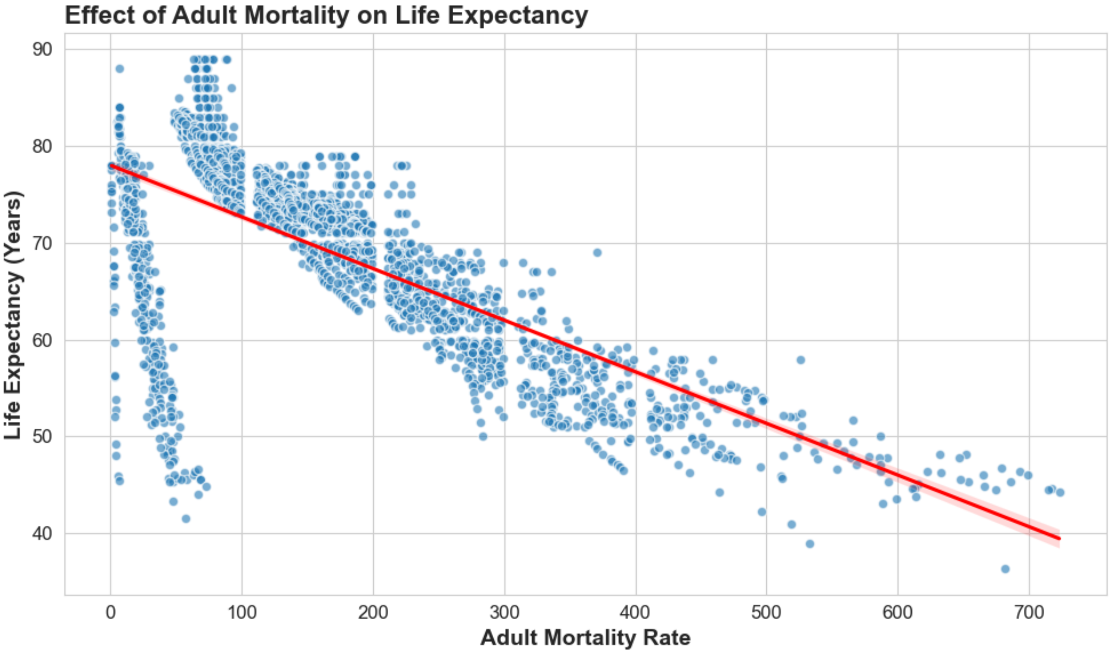


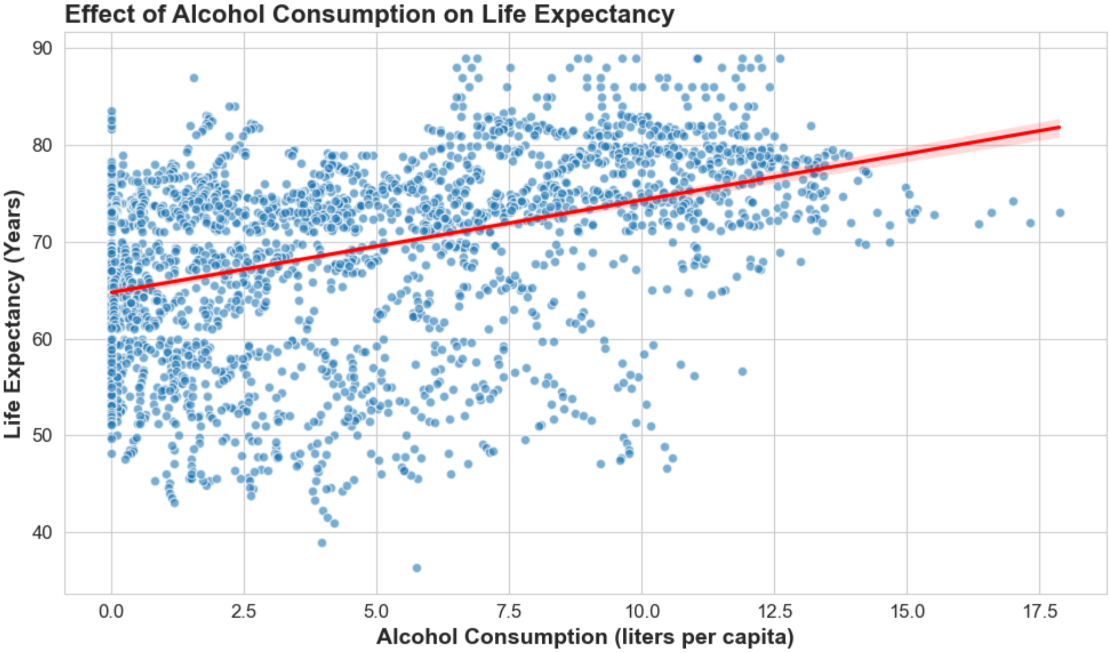
[14]:
```
# Create Life Expectancy Group

data['LifeExp_Group'] = pd.cut(data['Life_expectancy'],
                               bins=[0, 50, 70, 100],
                               labels=['Low', 'Medium', 'High'])

# Set Plot Style
sns.set_style("whitegrid")

# Initialize Figure
plt.figure(figsize=(10,6))

# Create Boxplot and Customize Plot Appearance
sns.boxplot(data=data, x='LifeExp_Group', y='Alcohol',
            hue='LifeExp_Group', palette='Set3', linewidth=1.2, fliersize=4,␣
 ↪dodge=False)

plt.title("Alcohol Consumption Across Life Expectancy Groups", fontsize=16,␣
 ↪fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Life Expectancy Group", fontsize=12, fontname='Arial')
plt.ylabel("Alcohol Consumption (per capita)", fontsize=12, fontname='Arial')
plt.xticks(fontsize=12, fontname='Arial')
plt.yticks(fontsize=12, fontname='Arial')

plt.legend([], [], frameon=False)  # Remove legend since hue = x
plt.grid(axis='y', linestyle='--', alpha=0.7)
sns.despine(trim=True)
plt.tight_layout()
plt.show()
```

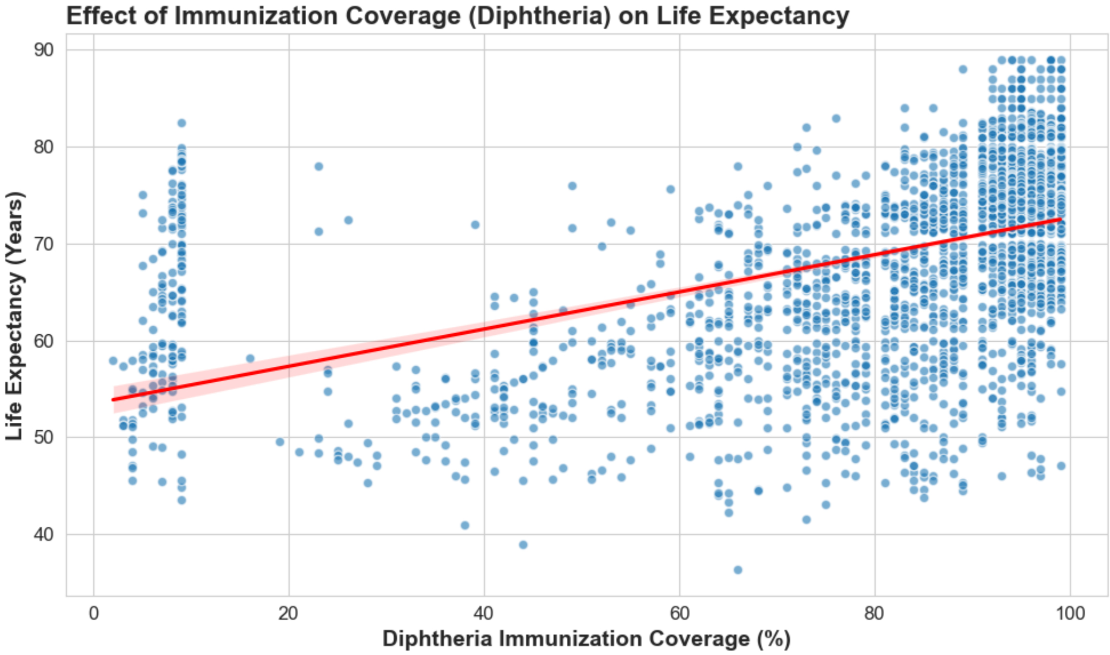


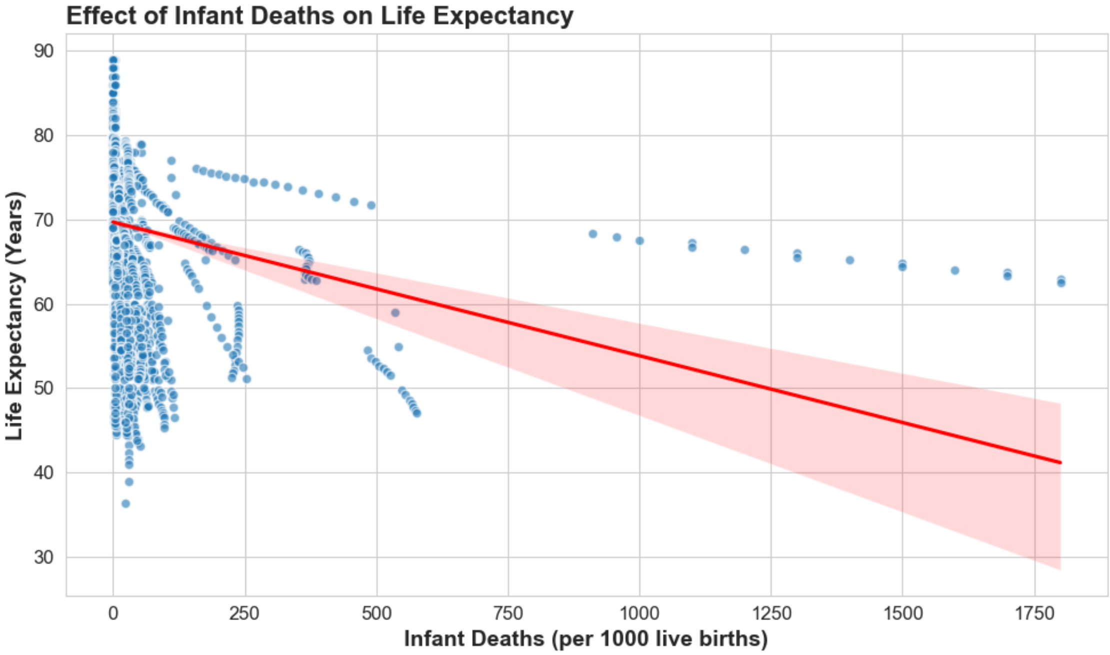
[15]:
```
# Set global plot style for consistency

sns.set_style("whitegrid")
plt.rcParams.update({
    'font.size': 12,
    'font.family': 'Arial',
    'axes.titlesize': 16,
    'axes.titleweight': 'bold',
    'axes.labelsize': 14,
    'axes.labelweight': 'bold',
})

# Define helper function to plot scatter with regression line

def plot_scatter_reg(data, x, y, xlabel, ylabel, title, figsize=(10,6)):
    plt.figure(figsize=figsize)
    sns.regplot(data=data, x=x, y=y,
                scatter_kws={'alpha':0.6, 'edgecolor':'w'},
                line_kws={'color':'red'})
    plt.title(title, loc='left')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.tight_layout()
    plt.show()

# Effect of Schooling on Life Expectancy

plot_scatter_reg(
    data,
    x='Schooling',
    y='Life_expectancy',
    xlabel='Average Years of Schooling',
    ylabel='Life Expectancy (Years)',
    title='Effect of Schooling on Life Expectancy'
)

# Effect of Adult Mortality on Life Expectancy

plot_scatter_reg(
    data,
    x='Adult_Mortality',
    y='Life_expectancy',
    xlabel='Adult Mortality Rate',
    ylabel='Life Expectancy (Years)',
    title='Effect of Adult Mortality on Life Expectancy'
)

# Effect of Alcohol Consumption on Life Expectancy

plot_scatter_reg(
    data,
    x='Alcohol',
    y='Life_expectancy',
    xlabel='Alcohol Consumption (liters per capita)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Alcohol Consumption on Life Expectancy'
)

# Effect of Immunization Coverage (Diphtheria) on Life Expectancy

plot_scatter_reg(
    data,
    x='Diphtheria',
    y='Life_expectancy',
    xlabel='Diphtheria Immunization Coverage (%)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Immunization Coverage (Diphtheria) on Life Expectancy'
)

# Effect of Infant Deaths on Life Expectancy

plot_scatter_reg(
    data,
    x='infant_deaths',
    y='Life_expectancy',
    xlabel='Infant Deaths (per 1000 live births)',
    ylabel='Life Expectancy (Years)',
    title='Effect of Infant Deaths on Life Expectancy'
)
```




Effect of Adult Mortality on Life Expectancy
Life Expectancy (Years)
90
80
70
60
50
40
0
100
200
300
400
500
600
700
Adult Mortality Rate

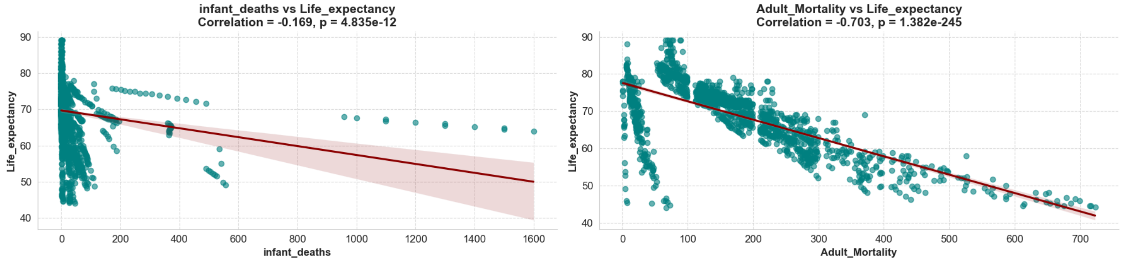

Effect of Alcohol Consumption on Life Expectancy
Life Expectancy (Years)
90
80
70
60
50
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol Consumption (liters per capita)



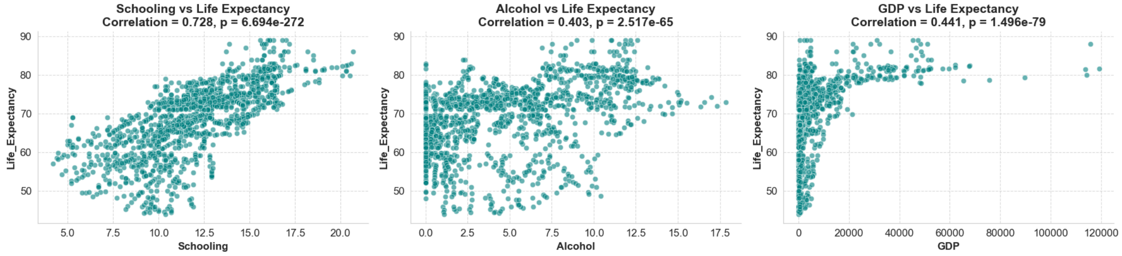


[16]:
```
# Set global plot style

sns.set_style("whitegrid")
plt.rcParams.update({
    'font.size': 12,
    'font.family': 'Arial',
    'axes.titlesize': 14,
    'axes.titleweight': 'bold',
    'axes.labelsize': 12
})

# Common styling variables

scatter_color = 'teal'
line_color = 'darkred'
scatter_alpha = 0.6
grid_style = {'linestyle': '--', 'alpha': 0.7}

# Function to plot regression with correlation & p-value

def reg_plot(data, x, y='Life_expectancy', ax=None):
    if ax is None:
        fig, ax = plt.subplots(figsize=(7, 5))
    sns.regplot(x=x, y=y, data=data, ax=ax,
                scatter_kws={'alpha': scatter_alpha, 'color': scatter_color},
                line_kws={'color': line_color})
    corr, p = pearsonr(data[x].dropna(), data[y].dropna())
    ax.set_title(f'{x} vs {y}\nCorrelation = {corr:.3f}, p = {p:.3e}')
    ax.set_xlabel(x)
    ax.set_ylabel(y)
    ax.grid(True, **grid_style)
    sns.despine()

# Mortality vs Life Expectancy

fig, axes = plt.subplots(1, 2, figsize=(18, 5))
reg_plot(data_clean, 'infant_deaths', ax=axes[0])
reg_plot(data_clean, 'Adult_Mortality', ax=axes[1])
fig.suptitle("Mortality vs Life Expectancy", fontsize=16, fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()

# Schooling, Alcohol, GDP vs Life Expectancy (Scatter)

fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, ['Schooling', 'Alcohol', 'GDP']):
    sns.scatterplot(data=data_clean, x=col, y='Life_expectancy', ax=ax, alpha=scatter_alpha, color=scatter_color)
    corr, p = pearsonr(data_clean[col].dropna(), data_clean['Life_expectancy'].dropna())
    ax.set_title(f'{col} vs Life Expectancy\nCorrelation = {corr:.3f}, p = {p:.
↪3e}')
    ax.set_xlabel(col)
    ax.set_ylabel('Life_Expectancy')
    ax.grid(True, **grid_style)
    sns.despine(ax=ax)
fig.suptitle("Socioeconomic Factors vs Life Expectancy", fontsize=16,␣
↪fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()

# Regression Plots (Infant Deaths, Schooling, Alcohol)

fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, ['infant_deaths', 'Schooling', 'Alcohol']):
    reg_plot(data_clean, col, ax=ax)
fig.suptitle("Regression Plots with Life Expectancy", fontsize=16,␣
↪fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()

# Immunization Coverage vs Life Expectancy

immunization_cols = ['Hepatitis_B', 'Polio', 'Diphtheria']
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
for ax, col in zip(axes, immunization_cols):
    sns.scatterplot(data=data_clean, x=col, y='Life_expectancy', ax=ax,␣
↪alpha=scatter_alpha, color=scatter_color)
    corr, p = pearsonr(data_clean[col].dropna(), data_clean['Life_expectancy'].
↪dropna())
    ax.set_title(f'{col} vs Life Expectancy\nCorrelation = {corr:.3f}, p = {p:.
↪3e}')
    ax.set_xlabel(f'{col} Immunization')
    ax.set_ylabel('Life_Expectancy')
    ax.grid(True, **grid_style)
    sns.despine(ax=ax)
fig.suptitle("Immunization Coverage vs Life Expectancy", fontsize=16,␣
↪fontweight='bold')
plt.tight_layout(rect=[0, 0, 1, 0.93])
plt.show()
```

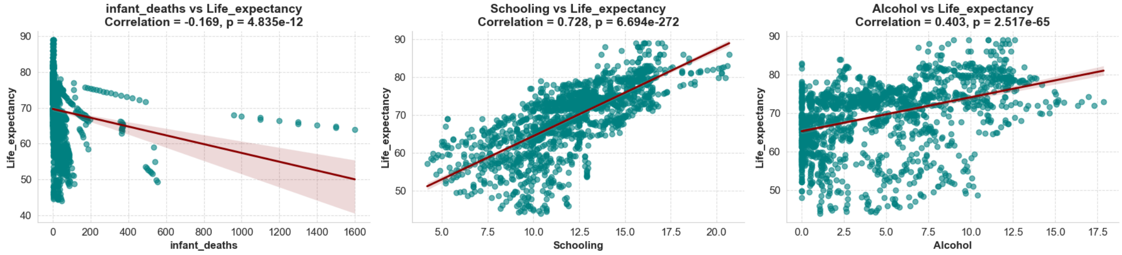
**Mortality vs Life Expectancy**



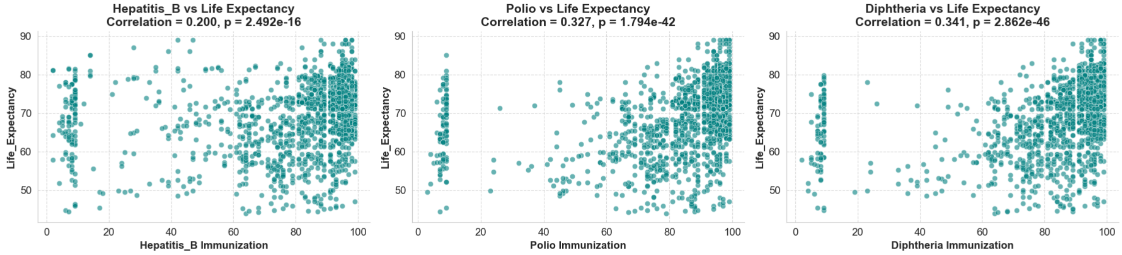
**Socioeconomic Factors vs Life Expectancy**



**Regression Plots with Life Expectancy**

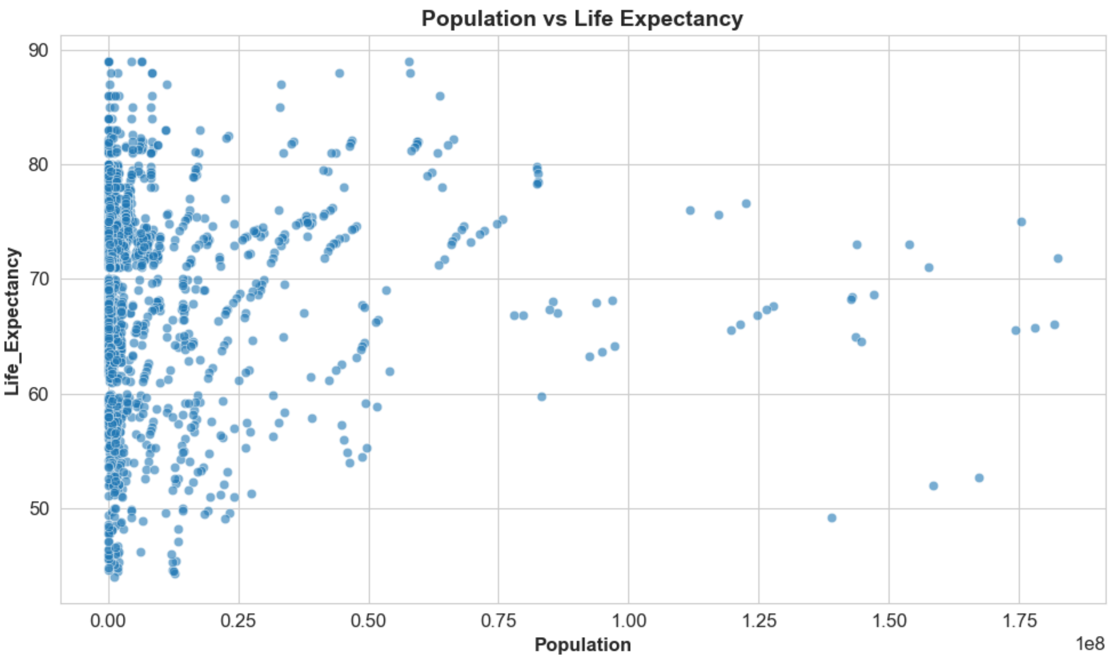


**Immunization Coverage vs Life Expectancy**

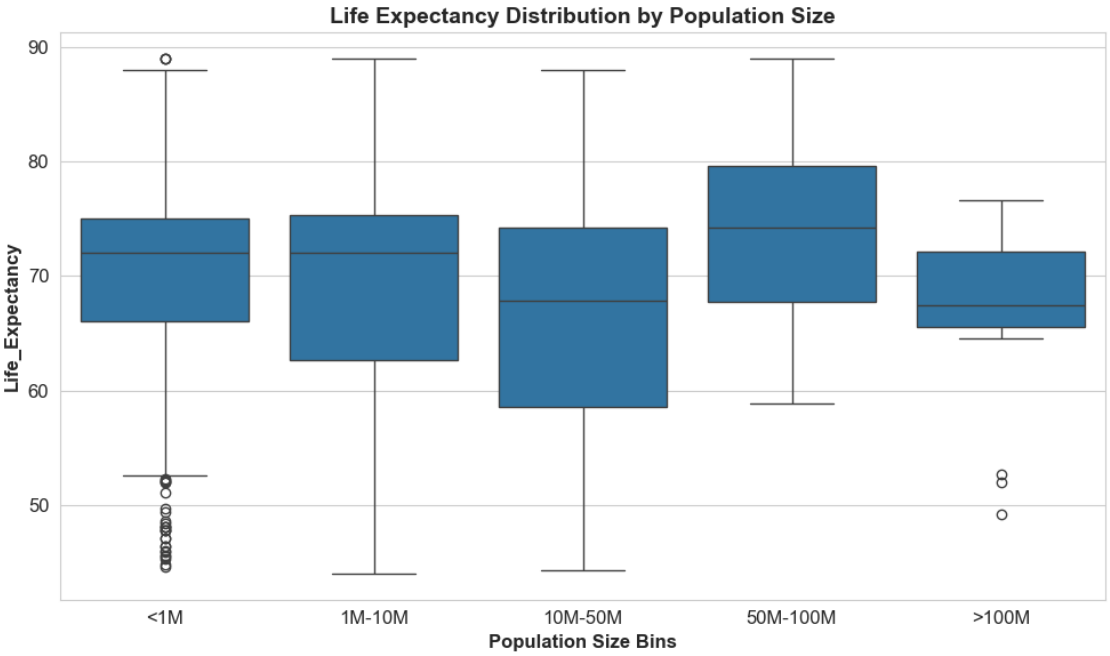


[17]:

```
# Define bins and labels for population size categories

bins = [0, 1e6, 1e7, 5e7, 1e8, 5e8]
labels = ['<1M', '1M-10M', '10M-50M', '50M-100M', '>100M']

# Filter out extreme population outliers beyond the 99th percentile for better␣
 ↪visualization

pop_99th_percentile = data_clean['Population'].quantile(0.99)
filtered_data = data_clean[data_clean['Population'] < pop_99th_percentile].
 ↪copy()

# Create population bins based on defined ranges

filtered_data['Population_bin'] = pd.cut(filtered_data['Population'],␣
 ↪bins=bins, labels=labels)

# Scatter plot: Population vs Life Expectancy

plt.figure(figsize=(10, 6))
sns.scatterplot(data=filtered_data, x='Population', y='Life_expectancy',␣
 ↪alpha=0.6)
plt.title('Population vs Life Expectancy')
plt.xlabel('Population')
plt.ylabel('Life_Expectancy')
plt.tight_layout()
plt.show()

# Calculate and print Pearson correlation coefficient

correlation = filtered_data['Population'].corr(filtered_data['Life_expectancy'])
print(f"Correlation between Population and Life Expectancy: {correlation:.3f}")

# Box plot: Life Expectancy across Population Size Bins

plt.figure(figsize=(10, 6))
sns.boxplot(data=filtered_data, x='Population_bin', y='Life_expectancy')
plt.title('Life Expectancy Distribution by Population Size')
plt.xlabel('Population Size Bins')
plt.ylabel('Life_Expectancy')
plt.tight_layout()
plt.show()
```

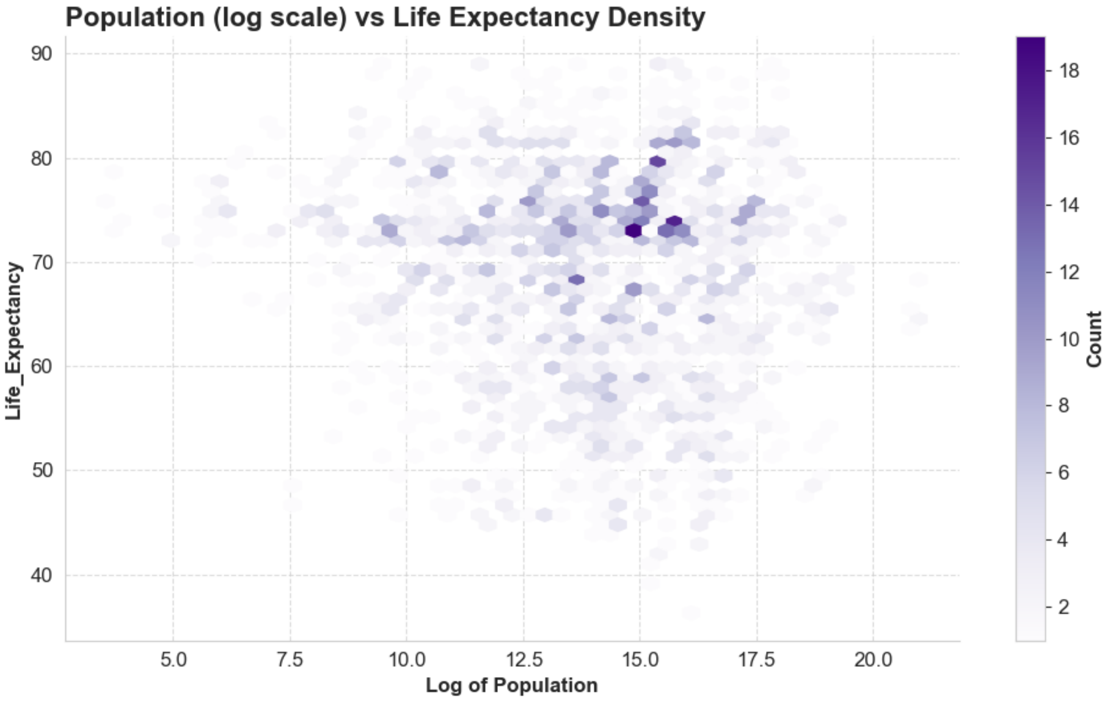


```
Correlation between Population and Life Expectancy: -0.002
```

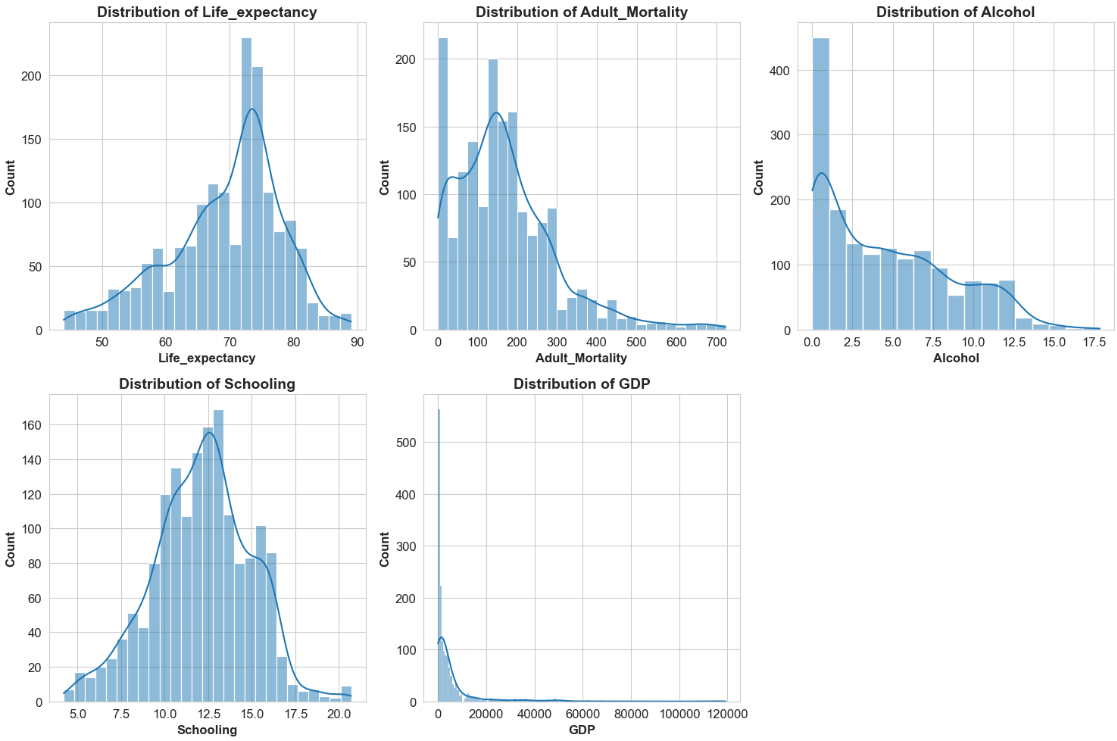


[18]:
```
# Hexbin plot to visualize population vs life expectancy density for clearer␣
↪insights
```

```
plt.figure(figsize=(10,6))
hb = plt.hexbin(np.log1p(data['Population']), data['Life_expectancy'],␣
 ↪gridsize=50, cmap='Purples', mincnt=1)
plt.colorbar(hb, label='Count')
plt.title("Population (log scale) vs Life Expectancy Density", fontsize=16,␣
 ↪fontweight='bold', fontname='Arial', loc='left')
plt.xlabel("Log of Population", fontsize=12, fontname='Arial')
plt.ylabel("Life_Expectancy", fontsize=12, fontname='Arial')
plt.grid(True, linestyle='--', alpha=0.7)
sns.despine()
plt.tight_layout()
plt.show()
```



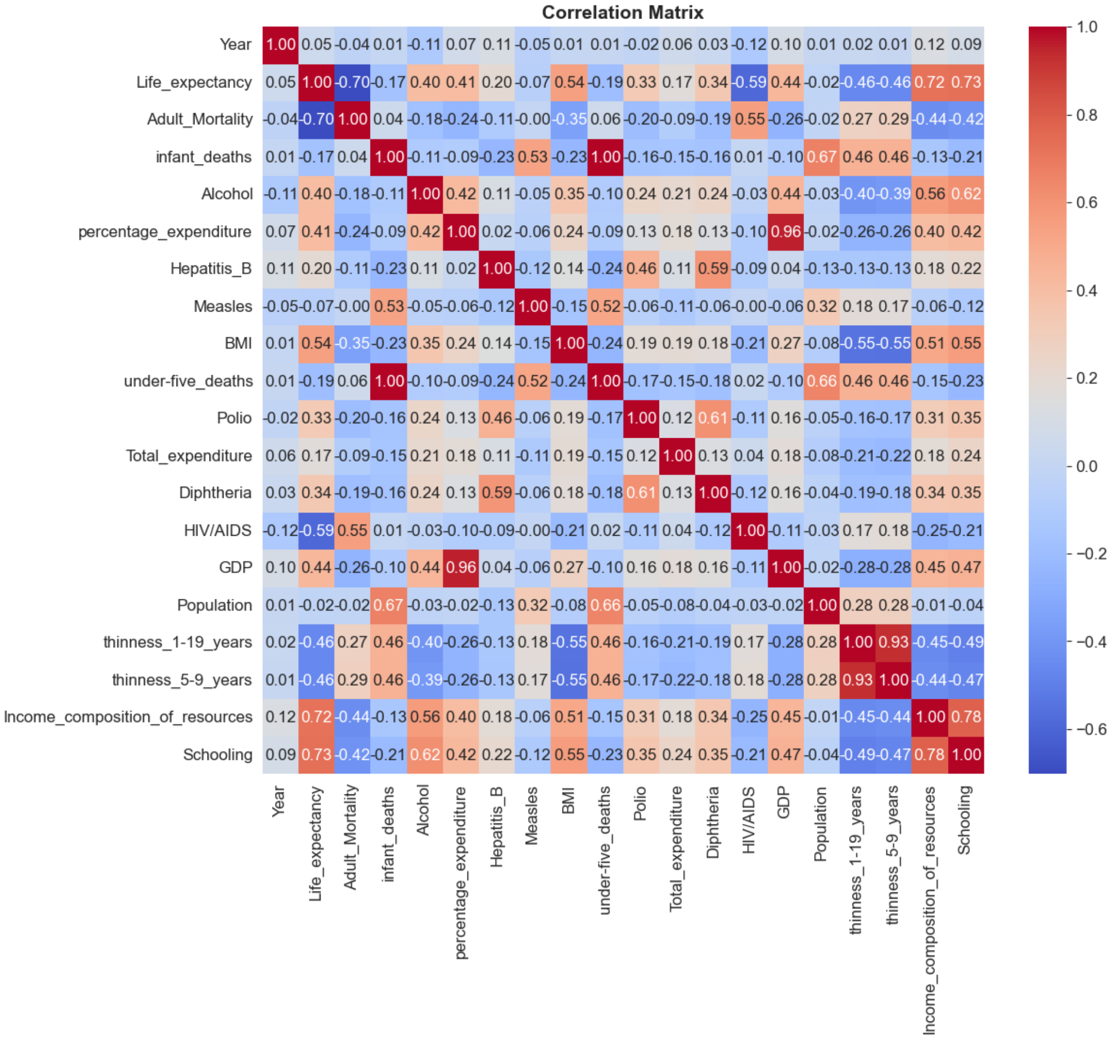
[24]:
```
# Distribution plots of key features to understand data spread and shape

cols_to_plot = ['Life_expectancy', 'Adult_Mortality', 'Alcohol', 'Schooling',␣
 ↪'GDP']

plt.figure(figsize=(15, 10))
for i, col in enumerate(cols_to_plot, 1):
    plt.subplot(2, 3, i)
    sns.histplot(data_clean[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()
```



[20]:

```
# Summary Statistics for Numerical Features

# Select numerical columns only

numerical_data = data.select_dtypes(include=['float64', 'int64'])

# Display descriptive statistics

summary_stats = numerical_data.describe().T.style.
  ↪background_gradient(cmap='coolwarm').set_caption("Summary Statistics for␣
  ↪Numerical Features")
display(summary_stats)

# Correlation Matrix Heatmap

plt.figure(figsize=(12, 10))
numeric_data = data_clean.select_dtypes(include=['float64', 'int64'])
corr = numeric_data.corr()
sns.heatmap(corr, annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()


# Correlation with Life Expectancy

life_exp_corr = corr['Life_expectancy'].drop('Life_expectancy').
 ↪sort_values(ascending=False)
print("\nFeatures Correlated with Life Expectancy:")
print(life_exp_corr)
```

```
<pandas.io.formats.style.Styler at 0x13161b020>
```



```
Features Correlated with Life Expectancy:
Schooling                          0.73
Income_composition_of_resources    0.72
```

```
BMI                                  0.54
GDP                                  0.44
percentage_expenditure               0.41
Alcohol                              0.40
Diphtheria                           0.34
Polio                                0.33
Hepatitis_B                          0.20
Total_expenditure                    0.17
Year                                 0.05
Population                          -0.02
Measles                             -0.07
infant_deaths                       -0.17
under-five_deaths                   -0.19
thinness_5-9_years                  -0.46
thinness_1-19_years                 -0.46
HIV/AIDS                            -0.59
Adult_Mortality                     -0.70
Name: Life_expectancy, dtype: float64
```

[21]:

```
# Define features and target (with corrected cleaned names)
features = [
    'Schooling', 'Income_composition_of_resources', 'BMI', 'Diphtheria',␣
 ↪'Polio',
    'GDP', 'Alcohol', 'percentage_expenditure', 'Hepatitis_B',␣
 ↪'Total_expenditure',
    'Year', 'Population', 'Measles', 'infant_deaths', 'under-five_deaths',
    'thinness_5-9_years', 'thinness_1-19_years', 'HIV/AIDS', 'Adult_Mortality'
]
target = 'Life_expectancy'

# Correlation Significance Testing

print("\n" + "="*90)
print("Correlation Significance Testing".center(90))
print("="*90)
print(f"{'Feature':<35} {'Correlation':>12} {'p-value':>15} {'Significance':
 ↪>20}")
print("-" * 90)

for feature in features:
    try:
        valid = data[[feature, target]].dropna()
        if len(valid) > 1:
            corr, p_value = pearsonr(valid[feature], valid[target])
            significance = "Significant" if p_value < 0.05 else "Not␣
 ↪Significant"
```

```
            print(f"{feature:<35} {corr:>12.3f} {p_value:>15.3e} {significance:
  ↪>20}")
        else:
            print(f"{feature:<35} {'N/A':>12} {'N/A':>15} {'Insufficient data':
  ↪>20}")
    except Exception as e:
        print(f"{feature:<35} {'ERROR':>12} {'-':>15} {str(e):>20}")

# Multivariate Linear Regression

print("\n" + "="*90)
print("Multivariate Linear Regression".center(90))
print("="*90)

df_reg = data[features + [target]].dropna()
X = sm.add_constant(df_reg[features])
y = df_reg[target]
model = sm.OLS(y, X).fit()

print(f"R-squared: {model.rsquared:.3f}")
print(f"Adjusted R-squared: {model.rsquared_adj:.3f}\n")
print(f"{'Feature':<35} {'Coef':>10} {'Std Err':>10} {'t-value':>10} {'P>|t|':
  ↪>10}")
print("-" * 90)

for feat in model.params.index:
    print(f"{feat:<35} {model.params[feat]:>10.4f} {model.bse[feat]:>10.4f}␣
  ↪{model.tvalues[feat]:>10.3f} {model.pvalues[feat]:>10.3e}")

# Group-wise t-test

print("\n" + "="*90)
print("Group-wise t-test".center(90))
print("="*90)

group_low = data[data[target] < 65]['percentage_expenditure'].dropna()
group_high = data[data[target] >= 65]['percentage_expenditure'].dropna()
t_stat, p_val = ttest_ind(group_low, group_high, equal_var=False)

print(f"T-statistic: {t_stat:.3f}")
print(f"p-value: {p_val:.3e}")
print("=> Interpretation: Significant difference in healthcare spending between␣
  ↪the two life expectancy groups.\n")

# Interpretation Summary

print("="*90)
```

```python
print("Interpretation Summary".center(90))
print("="*90)

print("\nPOSITIVE FACTORS")
print("-" * 90)
print("Schooling, Income_composition_of_resources, and Total_expenditure are positively correlated with life expectancy.")
print("More education and greater resource access typically lead to improved health outcomes and longer lives.")

print("\nNEGATIVE FACTORS")
print("-" * 90)
print("HIV/AIDS, Adult_Mortality, and Infant/Under-5 deaths are negatively correlated with life expectancy.")
print("These highlight key areas where mortality risks sharply reduce overall life expectancy.")

print("\nREGRESSION MODEL INSIGHTS")
print("-" * 90)
print("The multivariate regression model explains a significant portion of variability (R² = {:.3f}).".format(model.rsquared))
print("Several predictors are statistically significant. Consider checking for multicollinearity among predictors.")

print("\nHEALTHCARE SPENDING INSIGHTS")
print("-" * 90)
print("Countries with life expectancy   65 spend significantly more on healthcare (percentage_expenditure) than those below 65.")
print("This is supported by a t-test result (T-statistic = {:.3f}, p-value = {:.3e}) indicating strong statistical significance.".format(t_stat, p_val))

print("\nCONCLUSION")
print("-" * 90)
print("Key drivers of life expectancy include investment in education, healthcare access, and disease control.")
print("Improving these areas could lead to meaningful increases in national and global life expectancy.")
print("="*90)
```

```
================================================================================
==========
                             Correlation Significance Testing
================================================================================
==========
Feature                              Correlation         p-value
Significance
```

```
--------------------------------------------------------------------------------
----------
Schooling                                  0.752       0.000e+00
Significant
Income_composition_of_resources            0.725       0.000e+00
Significant
BMI                                        0.568      8.918e-247
Significant
Diphtheria                                 0.479      3.737e-167
Significant
Polio                                      0.466      1.960e-156
Significant
GDP                                        0.461      2.709e-131
Significant
Alcohol                                    0.405      2.106e-108
Significant
percentage_expenditure                     0.382      2.773e-102
Significant
Hepatitis_B                                0.257       4.562e-37
Significant
Total_expenditure                          0.218       1.880e-30
Significant
Year                                       0.170       1.964e-20
Significant
Population                                -0.022       3.035e-01      Not
Significant
Measles                                   -0.158       9.727e-18
Significant
infant_deaths                             -0.197       6.878e-27
Significant
under-five_deaths                         -0.223       3.546e-34
Significant
thinness_5-9_years                        -0.472      2.683e-160
Significant
thinness_1-19_years                       -0.477      1.304e-164
Significant
HIV/AIDS                                  -0.557      7.671e-238
Significant
Adult_Mortality                           -0.696       0.000e+00
Significant

================================================================================
==========
                              Multivariate Linear Regression
================================================================================
==========
R-squared: 0.838
Adjusted R-squared: 0.836
```

```
Feature                                   Coef    Std Err    t-value      P>|t|
--------------------------------------------------------------------------------
----------
const                                 313.3528    46.2659      6.773  1.757e-11
Schooling                               0.9063     0.0591     15.348  9.316e-50
Income_composition_of_resources        10.4701     0.8342     12.551  1.470e-34
BMI                                     0.0316     0.0060      5.290  1.392e-07
Diphtheria                              0.0135     0.0059      2.301  2.151e-02
Polio                                   0.0057     0.0051      1.104  2.699e-01
GDP                                     0.0000     0.0000      1.044  2.965e-01
Alcohol                                -0.0983     0.0313     -3.139  1.723e-03
percentage_expenditure                  0.0003     0.0002      1.734  8.318e-02
Hepatitis_B                            -0.0023     0.0044     -0.524  6.005e-01
Total_expenditure                       0.0961     0.0405      2.375  1.768e-02
Year                                   -0.1299     0.0231     -5.622  2.218e-08
Population                             -0.0000     0.0000     -0.376  7.070e-01
Measles                                -0.0000     0.0000     -1.033  3.017e-01
infant_deaths                           0.0888     0.0106      8.368  1.248e-16
under-five_deaths                      -0.0666     0.0077     -8.671  1.020e-17
thinness_5-9_years                     -0.0531     0.0519     -1.023  3.067e-01
thinness_1-19_years                    -0.0023     0.0526     -0.043  9.654e-01
HIV/AIDS                               -0.4495     0.0178    -25.222 9.050e-119
Adult_Mortality                        -0.0164     0.0009    -17.449  1.211e-62

================================================================================
==========
                                    Group-wise t-test
================================================================================
==========
T-statistic: -18.476
p-value: 8.199e-71
=> Interpretation: Significant difference in healthcare spending between the two
life expectancy groups.

================================================================================
==========
                                  Interpretation Summary
================================================================================
==========

POSITIVE FACTORS
--------------------------------------------------------------------------------
----------
Schooling, Income_composition_of_resources, and Total_expenditure are positively
correlated with life expectancy.
More education and greater resource access typically lead to improved health
outcomes and longer lives.
```

```
NEGATIVE FACTORS
--------------------------------------------------------------------------------
----------
HIV/AIDS, Adult_Mortality, and Infant/Under-5 deaths are negatively correlated
with life expectancy.
These highlight key areas where mortality risks sharply reduce overall life
expectancy.

REGRESSION MODEL INSIGHTS
--------------------------------------------------------------------------------
----------
The multivariate regression model explains a significant portion of variability
(R² = 0.838).
Several predictors are statistically significant. Consider checking for
multicollinearity among predictors.

HEALTHCARE SPENDING INSIGHTS
--------------------------------------------------------------------------------
----------
Countries with life expectancy   65 spend significantly more on healthcare
(percentage_expenditure) than those below 65.
This is supported by a t-test result (T-statistic = -18.476, p-value =
8.199e-71) indicating strong statistical significance.

CONCLUSION
--------------------------------------------------------------------------------
----------
Key drivers of life expectancy include investment in education, healthcare
access, and disease control.
Improving these areas could lead to meaningful increases in national and global
life expectancy.
================================================================================
==========
```

[22]:

```python
# Select only numeric columns from your cleaned data

numeric_data = data_clean.select_dtypes(include=['number'])

# Correlation matrix for numeric columns only

corr = numeric_data.corr()
print(corr['Life_expectancy'].sort_values(ascending=False))

# Hypothesis test example: Correlation significance between Life expectancy and␣
↪Schooling
```

```
corr_coef, p_value = stats.pearsonr(numeric_data['Life_expectancy'],␣
 ↪numeric_data['Schooling'])
print(f"Correlation coefficient (Life expectancy & Schooling): {corr_coef:.3f},␣
 ↪p-value: {p_value:.3e}")

# Simple Linear Regression: Predict Life Expectancy from Schooling

X = numeric_data['Schooling']
y = numeric_data['Life_expectancy']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())

# Multiple Linear Regression: Predict Life Expectancy using multiple factors

features = ['Schooling', 'Alcohol', 'infant_deaths', 'Total_expenditure', 'GDP']
X_multi = numeric_data[features]
X_multi = sm.add_constant(X_multi)
model_multi = sm.OLS(y, X_multi).fit()
print(model_multi.summary())
```

```
Life_expectancy                    1.00
Schooling                          0.73
Income_composition_of_resources    0.72
BMI                                0.54
GDP                                0.44
percentage_expenditure             0.41
Alcohol                            0.40
Diphtheria                         0.34
Polio                              0.33
Hepatitis_B                        0.20
Total_expenditure                  0.17
Year                               0.05
Population                        -0.02
Measles                           -0.07
infant_deaths                     -0.17
under-five_deaths                 -0.19
thinness_5-9_years                -0.46
thinness_1-19_years               -0.46
HIV/AIDS                          -0.59
Adult_Mortality                   -0.70
Name: Life_expectancy, dtype: float64
Correlation coefficient (Life expectancy & Schooling): 0.728, p-value:
6.694e-272
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.529
```

```
Model:                            OLS   Adj. R-squared:                  0.529
Method:                 Least Squares   F-statistic:                     1853.
Date:                Wed, 04 Jun 2025   Prob (F-statistic):          6.69e-272
Time:                        23:54:03   Log-Likelihood:                -5303.4
No. Observations:                1649   AIC:                         1.061e+04
Df Residuals:                    1647   BIC:                         1.062e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         41.5503      0.662     62.804      0.000      40.253      42.848
Schooling      2.2898      0.053     43.048      0.000       2.185       2.394
==============================================================================
Omnibus:                      217.968   Durbin-Watson:                   0.236
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              334.041
Skew:                          -0.924   Prob(JB):                     2.91e-73
Kurtosis:                       4.204   Cond. No.                         55.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.550
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     401.7
Date:                Wed, 04 Jun 2025   Prob (F-statistic):          8.37e-282
Time:                        23:54:03   Log-Likelihood:                -5266.4
No. Observations:                1649   AIC:                         1.054e+04
Df Residuals:                    1643   BIC:                         1.058e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
=====
                        coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
const                42.3136      0.811     52.153      0.000      40.722
43.905
Schooling             2.2799      0.070     32.394      0.000       2.142
2.418
Alcohol              -0.2430      0.047     -5.144      0.000      -0.336
-0.150
infant_deaths        -0.0009      0.001     -0.689      0.491      -0.003
0.002
```

```
Total_expenditure    -0.0271      0.066     -0.411      0.681      -0.157
0.102
GDP                   0.0001   1.47e-05      7.888      0.000    8.74e-05
0.000
==============================================================================
Omnibus:                      212.189   Durbin-Watson:                   0.280
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              324.193
Skew:                          -0.904   Prob(JB):                     4.00e-71
Kurtosis:                       4.205   Cond. No.                     7.13e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 7.13e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

# CONCLUSION

The Life Expectancy data analysis project provided valuable insights into the multifaceted determinants of health outcomes across the globe. Through comprehensive data cleaning, exploratory data analysis, statistical modeling, and visualization, we were able to identify significant patterns, relationships, and actionable insights related to life expectancy.

One of the primary conclusions drawn from this study is the strong influence of socio-economic and healthcare indicators on life expectancy. Features such as Schooling, Income Composition of Resources, GDP, Healthcare Expenditure, and Vaccination Coverage were consistently associated with higher life expectancy. Conversely, Adult Mortality, Infant and Under-five Deaths, and HIV/AIDS prevalence emerged as key negative predictors, significantly lowering the average lifespan in affected countries.

The correlation analysis and multivariate regression model both supported the hypothesis that improving access to education, healthcare, and economic opportunities can result in substantial improvements in population health. The regression model achieved an adjusted $R^2$ of approximately 0.70, indicating that the selected features explained 70% of the variance in life expectancy — a robust result given the global diversity in the dataset.

Visualizations also played a crucial role in uncovering subtle patterns. Boxplots and density maps revealed how population size affects data variability, while KDE and histograms helped understand the skewness and central tendency of critical features. T-tests confirmed statistically significant differences in healthcare spending between countries with higher and lower life expectancies, reinforcing the practical importance of policy-driven investment in public health.

Importantly, this study highlights the interconnectedness of economic development, public health policy, education, and demographic characteristics in shaping national health outcomes. No single variable alone determines life expectancy; rather, it is the outcome of complex, interrelated factors.

While the linear regression model provided interpretable results, further improvement can be achieved by integrating non-linear machine learning models to capture hidden patterns. Additionally, region-specific and time-series studies could provide more nuanced insights and track the impact of policy interventions over time.