PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |
| --- |
| TITLE: Life Expectancy |
| NAME: Guna Seelan C |
| COURSE: DA/DS, Offline |
| BATCH: June 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| 1. | **Introduction** |
| 2. | **Aim** |
| 3. | **Business Problem / Problem Statement** |
| 4. | **Project Workflow** |
| 5. | **Data Understanding** |
| 6. | **Data Cleaning**  Missing Values Imputation  Outlier Treatment  Handling Inconsistent Values |
| 7. | **Obtaining Derived Metrics** |
| 8. | **Filtering Data for Analysis** |
| 9. | **Statistical Analysis**  Descriptive analysis  Test statistics and hypothesis testing |
| 10. | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| 11. | **Bivariate Analysis** |
| 12. | **Multivariate Analysis** |
| 13. | **Overall Insights from Analysis** |
| 14. | **Conclusion** |

1. **INTRODUCTION**

This dataset provides a comprehensive view of global health indicators across multiple countries and years, with a focus on life expectancy and its influencing factors. It includes both developing and developed nations, allowing comparative analysis of health outcomes and socio-economic conditions.

1. **AIM OF THE PROJECT**

* To study factors that affect life expectancy in different countries
* To compare life expectancy between developed and developing nations
* To find trends in life expectancy over the years (2000–2015)
* To check how health spending and immunization impact life expectancy
* To build a model that can predict life expectancy using other variables
* To give useful suggestions for improving public health policies

1. **PROBLEM STATEMENT**

Life expectancy is different in each country. Some countries have people living longer, while others have shorter lifespans. Even though many governments spend money on healthcare and give vaccines, life expectancy is still low in some places. This project looks at data from 2000 to 2015 to find out what factors affect life expectancy—like income, education, diseases, and health services. The goal is to understand these patterns and help improve health policies for better living.

1. **PROJECT WORKFLOW**
2. **Data Collection**
   * Get the life expectancy dataset (2000–2015)
   * Check for missing values and data types
3. **Data Cleaning & Preprocessing**
   * Handle missing values
   * Convert data types if needed
   * Remove duplicates or irrelevant columns
4. **Exploratory Data Analysis (EDA)**
   * Use charts to study trends over time
   * Compare life expectancy across countries
   * Check relationships between variables (e.g., GDP vs. life expectancy)
5. **Visualization & Insights**
   * Identify key factors affecting life expectancy
   * Drop less useful columns
6. **Statistical Analysis**
   * Run correlation tests
   * Check significance of variable
7. **Model Building**
   * Train regression models to predict life expectancy
   * Evaluate model performance (R², RMSE, etc.)
8. **Insights & Recommendations**

* Summarize findings in simple language
* Suggest actions for improving health outcomes

1. **DATA UNDERSTANDING**

**Dataset Overview**

* **Rows:** ~2,900+ records
* **Columns:** 22
* **Each row represents a country in a specific year (2000–2015)**

**Key Columns**

* Country: Name of the country
* Year: Year of the record
* Status: Developed or Developing
* Life expectancy: Target variable (in years)

**Important Features**

* Adult Mortality: Death rate among adults
  + - Infant Deaths: Number of infant deaths per 1000 births
    - Alcohol: Alcohol consumption per capita
    - BMI: Average body mass index
    - GDP: Gross Domestic Product
    - Schooling: Average years of schooling
    - Income composition of resources: Income index
    - Hepatitis B, Polio, Diphtheria: Immunization rates

**Missing Values**

* Some columns like GDP, Alcohol, and BMI have missing values
  + Need to handle them before analysis

**Data Type**

* Mostly numerical columns
* A few categorical columns (Country, Status, Year)

**Target Variable**

* **Life expectancy is what we want to understand and predict**

1. **DATA CLEANING**

To ensure accurate analysis and meaningful insights, the dataset underwent a thorough cleaning process. This step corrected inconsistencies, handled missing values, and prepared the data for analysis.

1. **Missing Values Imputation**

Filling in missing data to maintain dataset completeness.

|  |  |
| --- | --- |
|  |  |

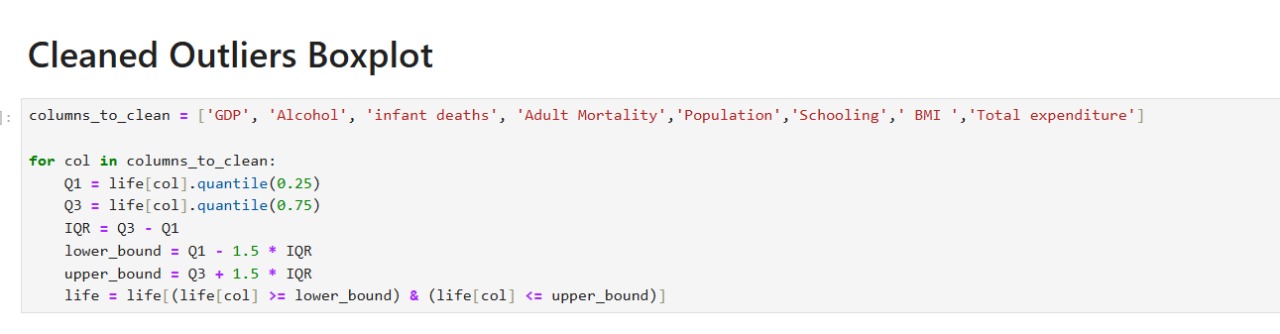


* Filled missing values in **Hepatitis B**, **Alcohol**, **Population**, **Total expenditure**, **Income composition of resources**, and **Schooling** using the **mean** of each column
* Filled missing values in **GDP** using the **median**, which is better for skewed data
* Converted the **Life expectancy** column to numeric format using pd.to\_numeric() to handle any non-numeric entries
* Ensured all columns are ready for analysis by handling missing data and fixing data types

1. **Outlier Treatment**

Extreme values were identified and removed or adjusted to prevent skewed results.

* Used Interquartile Range (IQR) method:



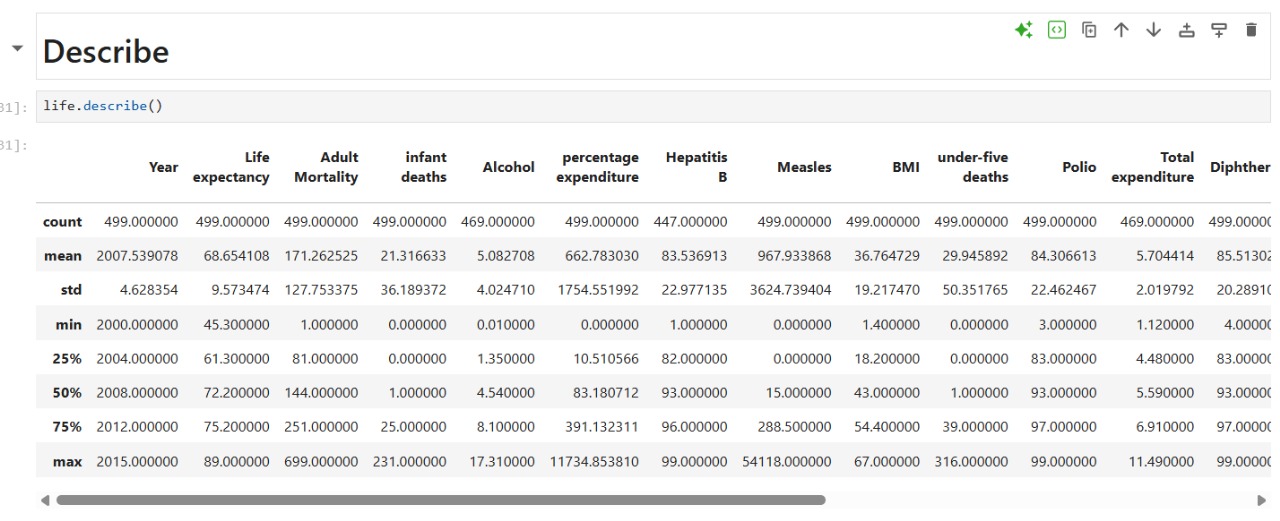
1. **OBTAINING DERIVED METRICES**

* Created new columns by performing calculations on existing data
* Example: Calculated **BMI** using weight and height if available (BMI = weight / height²)
* Derived **Health Index** by combining multiple health-related features
* Computed **Expenditure per capita** using total expenditure and population
* Normalized or scaled values to make comparisons easier across countries
* Used ratios (e.g., **GDP per capita**, **Alcohol per adult**) to get more meaningful insights
* Helped uncover hidden patterns and relationships in the data

1. **FILTERING DATA FOR ANALYSIS**

* **Removed rows with missing or invalid Life Expectancy values using dropna() or filtering after pd.to\_numeric()**
* **Selected only countries with Population > 1 million to focus on significant cases**
* **Filtered data for years after 2000 to analyze recent trends**
* **Excluded rows where GDP or Alcohol consumption was zero or extremely low (possible data entry errors)**
* **Focused on countries with Schooling > 10 years to study impact of education**
* **Used conditional filters like life[life['Life expectancy '] > 60] to analyze healthier populations**
* **Applied multiple filters together using logical operators (&, |) for refined subsets**

1. **STATISTICAL ANALYSIS**
   1. **Descriptive Analysis**

****

**2.Hypothesis testing**

Example Hypothesis 1:

**H₀ (Null Hypothesis):** Developed and developing countries have the same average life expectancy.  
**H₁ (Alternate Hypothesis):** Developed and developing countries have different average life expectancy

* **Test Used:** Independent two-sample t-test
* **Result:** p-value < 0.05 → Reject H₀  
  Developed countries show significantly higher life expectancy



Example Hypothesis 2:

**H₀:** Average life expectancy is the same across developed and developing countries.

* **Test Used:** One-way ANOVA
* **Result:** Significant variation in life expectancy between developed and developing countries



Example Hypothesis 3:

**H₀:** There is no difference in alcohol consumption between developed and developing countries.

* **Test Used:** Mann-Whitney U-Test
* • **Result:** No significant difference in alcohol consumption between the two groups (p-value = 0.080).

A screenshot of a computer

AI-generated content may be incorrect.

1. **EXPLORATORY DATA ANALYSIS (EDA)**

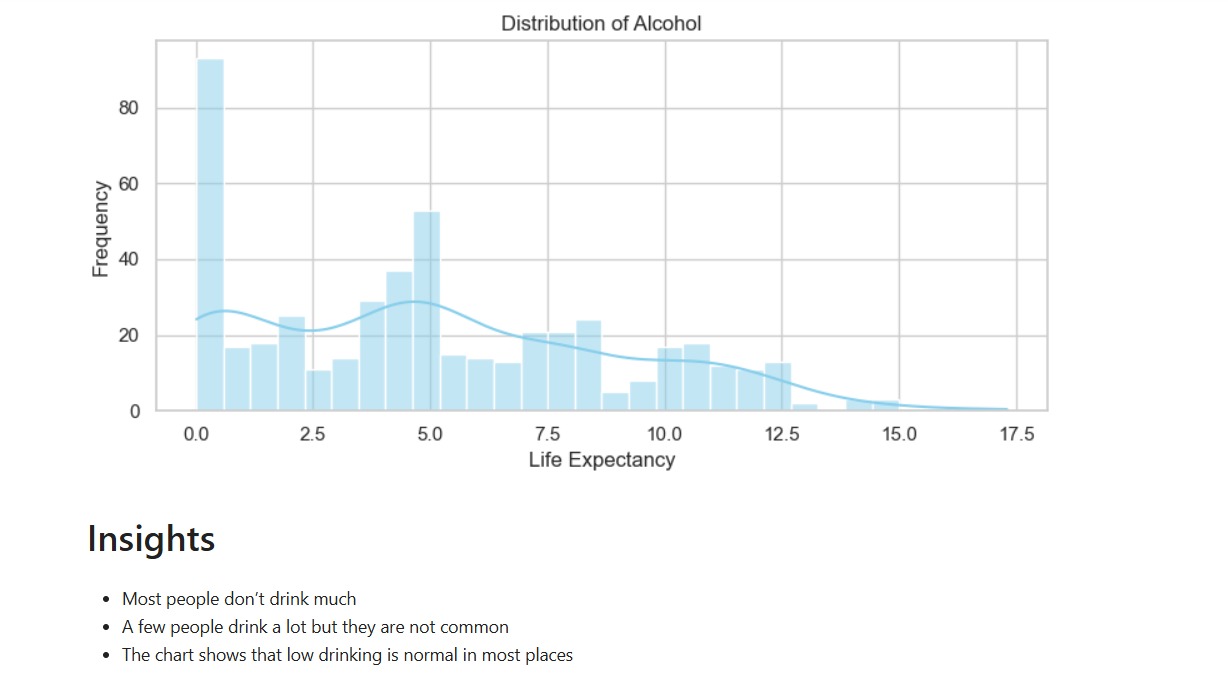
**UNIVARIATE ANALYSIS**

Univariate analysis focuses on exploring individual variables to understand their distribution, patterns, and anomalies.

| **Variable** | **Chart Type** | **Key Insights** |
| --- | --- | --- |
| Adult Mortality | Histogram + Density | Most countries fall between 100–300 deaths per 1000 adults; few extreme cases above 400. |
| Alcohol | Histogram + Line Graph | Low alcohol consumption is common; few countries show high usage. |
| BMI | Histogram | Most people are in normal or slightly overweight range; extremes are rare. |
| Life Expectancy | Histogram | Slight right skew; most countries have life expectancy between 60–75 years. |
| Schooling | Histogram | Majority of countries offer 10–15 years of schooling. |
| Health Expenditure | Histogram | Most countries spend less than 500 USD per capita; few high spenders. |
| GDP | Histogram | Most countries spend less than 500 USD per capita; few high spenders. |

A graph of a number of people

AI-generated content may be incorrect. A diagram of a distribution of bmi

AI-generated content may be incorrect.

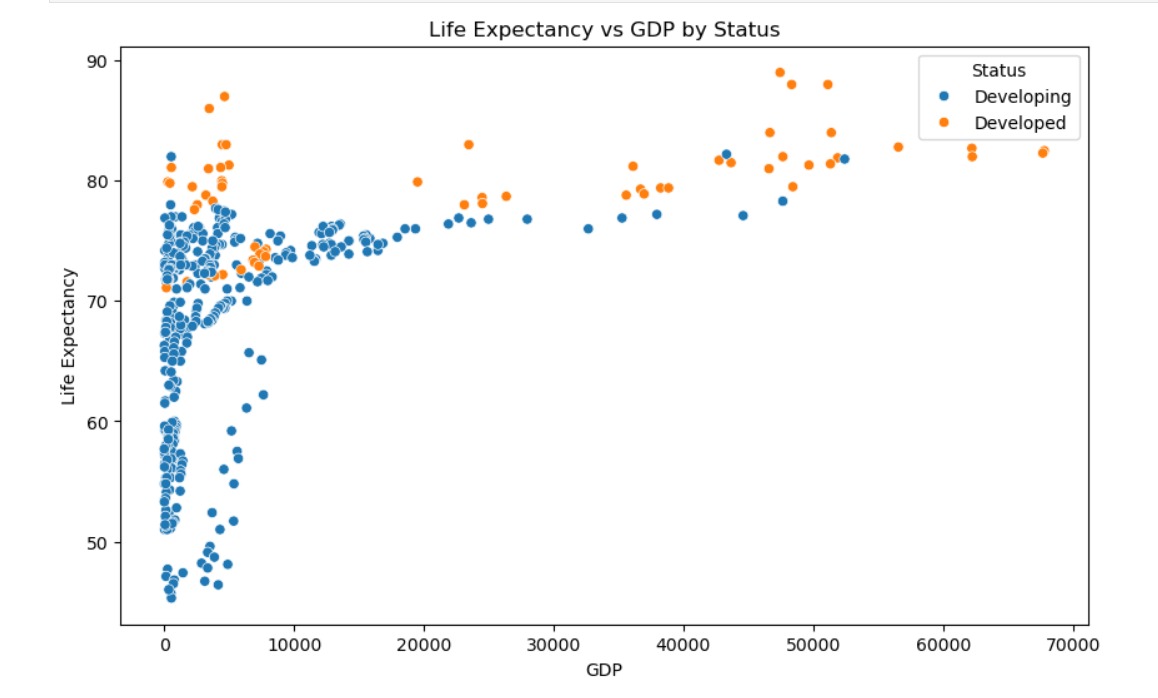
**Insights Gained:**

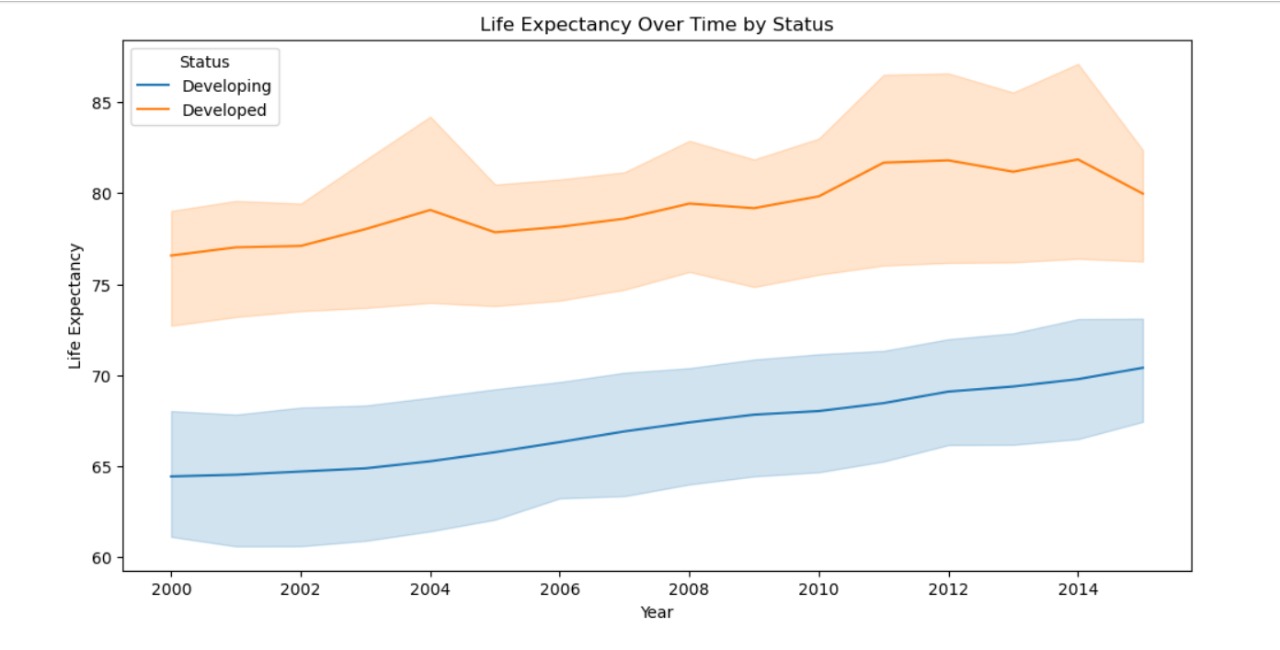
* Most countries have adult mortality between 100–300.
* Alcohol use is low in most places; few countries drink a lot.
* BMI is mostly in the normal to slightly high range.
* Life expectancy is usually between 60–75 years.

1. **BIVARIATE ANALYSIS**

Bivariate analysis helps explore relationships between two variables, uncovering trends, correlations, and disparities that influence life expectancy across countries.

| Variable Pair | Chart Type | Key Insights |
| --- | --- | --- |
| Life Expectancy vs GDP | Scatter Plot | Higher GDP is associated with longer life expectancy. |
| Status vs Life Expectancy | Line Graph | Developed countries consistently show higher life expectancy over time. |
| Year vs Life Expectancy | Line Graph | Life expectancy has increased steadily from 2000 to 2015 in both groups. |
| Status vs GDP | Scatter Plot | Developed countries generally have higher GDP than developing ones. |



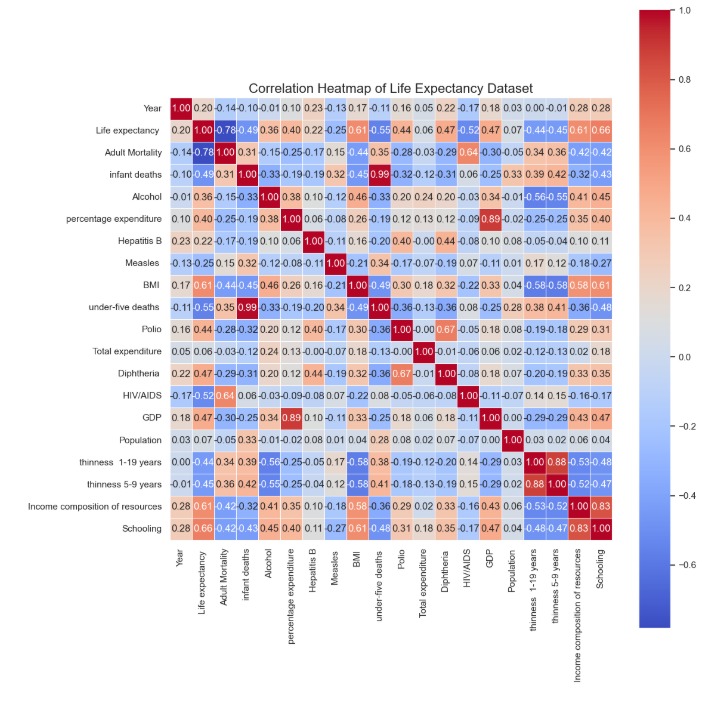
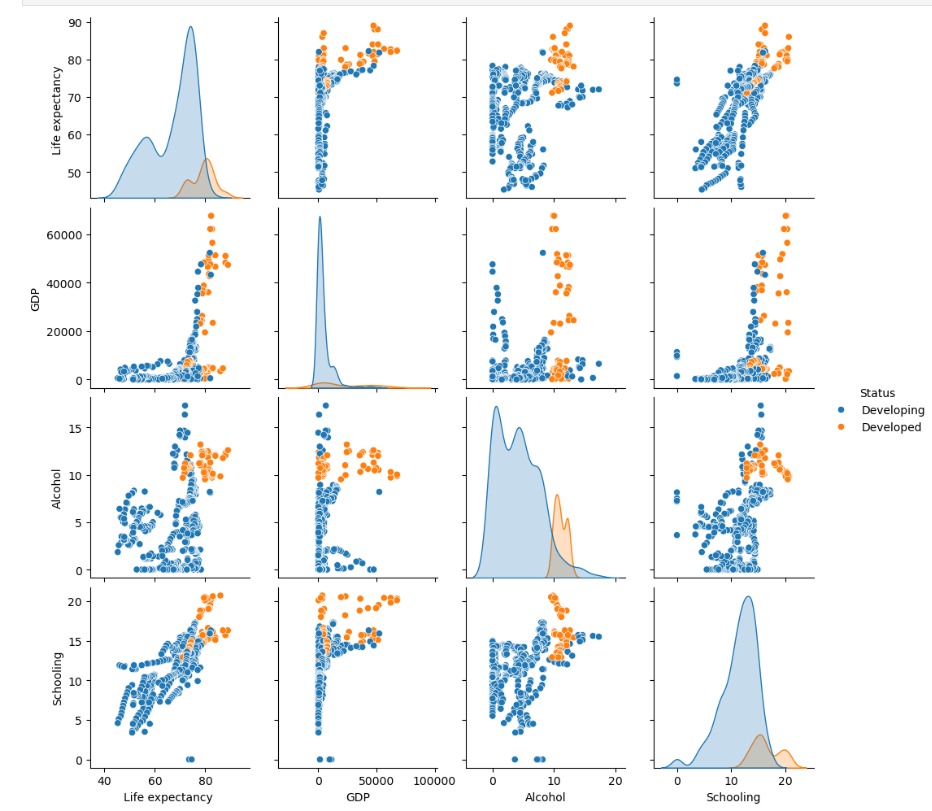


Key Takeaways:

* Life expectancy improves with higher GDP.
* Developed countries maintain a clear lead in life expectancy.
* Both groups show upward trends over time.
* Economic status strongly influences health outcomes

1. **MULTIVARAIATE ANALYSIS**

Multivariate analysis helps explore relationships between multiple variables at once, revealing deeper patterns and interactions that influence life expectancy across countries



**Key Takeaways:**

* **Schooling, GDP, and income composition** are strongly linked to higher life expectancy.
* **Adult mortality and HIV/AIDS** show strong negative correlations with life expectancy.
* Developed countries show tighter clusters with higher values across key health indicators.
* Multivariate patterns reveal that **economic and educational factors** work together to improve health outcomes.

1. **OVERALL INSIGHTS FROM ANALYSIS**
2. **Developed vs Developing Countries**
   * **Developed countries consistently show higher life expectancy, GDP, and schooling levels.**
3. **Key Positive Influencers**
   * **GDP, schooling, health expenditure, and immunization rates are positively correlated with life expectancy.**
4. **Negative Health Indicators**
   * **Adult mortality, HIV/AIDS prevalence, and infant deaths are strongly negatively correlated with life expectancy.**
5. **Trends Over Time**
   * **Life expectancy has steadily increased from 2000 to 2015 in both developed and developing countries, though the gap remains.**
6. **Alcohol and BMI**
   * Alcohol consumption and BMI show **mixed patterns**; most countries have moderate levels, with few outliers.
7. **Multivariate Patterns**
   * Economic and educational factors **work together** to improve health outcomes—countries with high GDP and schooling tend to have better life expectancy.
8. **Statistical Confirmation**
   * T-Test and ANOVA confirm **significant differences** in life expectancy between development statuses.
9. **Policy Implications**

* Improving education, healthcare access, and economic stability can **boost life expectancy**, especially in developing regions.

1. **CONCLUSION**

This life expectancy analysis project offered a comprehensive exploration of global health and development indicators, uncovering key patterns that influence how long people live across different countries.

**Key Conclusions:**

* **GDP per capita, schooling, and health expenditure** are strong predictors of life expectancy—higher values in these areas correlate with longer lifespans.
* **Immunization rates and BMI** also show positive associations with life expectancy, highlighting the importance of preventive care and nutrition.
* **Developed countries consistently exhibit higher life expectancy**, supported by statistical tests (T-Test, ANOVA) confirming significant group differences.
* **Alcohol consumption and HIV/AIDS prevalence** have complex relationships with life expectancy, requiring nuanced policy approaches.
* **Outliers and missing data** were identified and treated to ensure robust and reliable analysis.
* Derived metrics like **life expectancy gap, health spending ratio, and education-health synergy** added depth to the insights and enabled targeted recommendations.
* These findings can inform **policy makers and global health organizations** to prioritize education, healthcare access, and economic development in regions with lower life expectancy.