# Life Expectancy Analysis

## Data Analysis

### Project Report

Submitted by:

Anantha Krishnan K

BCE229

# 1. Project Introduction and Dataset Exploration

## 1.1 Introduction to the Project

Life expectancy is a key indicator of the overall health and development of a country. Analyzing the factors affecting life expectancy can provide crucial insights into global health trends, socio-economic conditions, and the impact of medical advancements.

In this project, we explore a dataset containing life expectancy data from multiple countries over the period **2000-2015**. Our goal is to analyze various health, economic, and social factors influencing life expectancy and derive meaningful insights.

## 1.2 Overview of the Dataset

The dataset consists of **2938 records** spanning **183 countries** over a **16-year period (2000-2015)**. It includes a mix of numerical and categorical features, covering economic, health, and demographic factors.

### 1.2.1 Dataset Features

The dataset contains the following key attributes:

| Feature | Description |
|---|---|
| Country | Name of the country |
| Year | Year of observation (2000-2015) |
| Status | Classification of the country as Developed or Developing |
| Life expectancy | Average expected lifespan at birth (Target Variable) |
| Adult Mortality | Probability of dying between ages 15-60 per 1000 population |
| Infant deaths | Number of infant deaths per 1000 births |
| Alcohol | Alcohol consumption per capita (liters) |
| Percentage expenditure | Expenditure on health as a percentage of GDP |
| Hepatitis B | Immunization rate for Hepatitis B in children (Percentage) |
| Measles | Number of measles cases reported |
| BMI | Average Body Mass Index |
| Polio | Immunization rate for Polio (Percentage) |
| Diphtheria | Immunization rate for Diphtheria (Percentage) |
| HIV/AIDS | Death rate due to HIV/AIDS per 1000 population |
| GDP | Gross Domestic Product per capita |
| Population | Total population of the country |
| Thinness 1-19 years | Percentage of underweight individuals aged 1-19 |
| Thinness 5-9 years | Percentage of underweight individuals aged 5-9 |
| Income composition of resources | Human Development Index (HDI) based measure of income |
| Schooling | Average number of years of schooling |

### 1.2.2 Initial Observations

- The dataset covers both developed and developing countries, allowing us to compare life expectancy trends between them.
- Several features relate to healthcare factors, such as immunization rates, mortality rates, and disease prevalence.
- Economic indicators like GDP and expenditure on health are included, which helps analyze their impact on life expectancy.
- Some columns contain missing values, requiring data cleaning before analysis.

- Outliers are present in features such as GDP, Population, and Measles cases, which will need treatment.

## 1.3 Objectives and Goals of the Analysis

The primary objective of this analysis is to understand the key factors influencing life expectancy and derive actionable insights.

### Key Goals:

- **Perform Data Cleaning**: Handle missing values, outliers, and inconsistencies.
- **Conduct Exploratory Data Analysis (EDA)**: Identify trends, patterns, and correlations in the data.
- **Apply Statistical Analysis**: Use hypothesis testing and statistical methods to validate insights.
- **Feature Selection and Model Building**: Determine the most important features affecting life expectancy.
- **Provide Data-Driven Insights**: Present conclusions that can help policymakers and health organizations improve life expectancy.

# 2. Data Cleaning Process

## 2.1 Handling Missing Values

During dataset exploration, we identified that some numerical features contained missing values. To ensure data integrity, we applied **median imputation**, as it is robust against outliers and does not distort the distribution.

### Strategy Used:

- **Numerical Features**: Replaced missing values with the **median** of the respective column.
- **Categorical Features**: No missing values were found, so no imputation was required.

## 2.2 Handling Outliers

Outliers can distort statistical analysis and negatively impact model performance. To address this, we applied three different strategies based on the severity of outliers in each feature:

### Outlier Detection Method:

We used the **Interquartile Range (IQR) method** to detect outliers:

### Outlier Handling Strategies:

1. **IQR Method** (For columns with **fewer than 50** outliers)
   a. Removed outliers that fell outside the 1.5*IQR range.
2. **Winsorization** (For columns with **50 to 300** outliers)
   a. Capped extreme values to the **5th and 95th percentiles** to maintain statistical properties while reducing extreme values.
3. **Log Transformation** (For columns with **more than 300** outliers)
   a. Applied **log transformation** using `np.log1p()` to normalize highly skewed features.

**Result**:

- IQR method **removed extreme outliers** in low-outlier features.
- Winsorization **capped extreme values** for moderately skewed features.
- Log transformation **normalized highly skewed features**.

## 2.4 Encoding Categorical Variables

The dataset contained **one categorical feature**:

- **Status**: ("Developed" / "Developing")

Since machine learning models require numerical inputs, we **converted this categorical feature into numerical values** using **Label Encoding**:

- "Developed" → **1**
- "Developing" → **0**

**Result:**
The "Status" column is now numerical, enabling it to be used in models.

## 2.3 Data Transformation (Scaling & Normalization)

Since the dataset contains variables with vastly different ranges (e.g., **GDP vs. Infant Deaths**), we applied **feature scaling** to standardize the dataset.

### Strategy Used:

1. **Standardization (Z-score transformation)** → Applied to features with different units or magnitudes

**Result:**

- **Standardization** transformed features to have a mean of 0 and a standard deviation of 1, making them comparable.

## Final Outcome of Data Cleaning

After completing the **data cleaning process**, we now have:

- **No missing values (handled using median imputation).**

- **Outliers detected and treated (IQR, Winsorization, Log Transformation).**
- **Categorical data converted into numerical format.**
- **Data transformation applied (Standardization).**

# 3. Exploratory Data Analysis (EDA)

## 3.1 Introduction to EDA

Exploratory Data Analysis (EDA) is a crucial step to understand the dataset's structure, detect patterns, and identify relationships between features. The primary objectives of EDA in this project are:

- Understanding the distribution of numerical features.
- Identifying correlations between features and the target variable (*Life Expectancy*).
- Detecting patterns or anomalies that could influence the analysis.
- Visualizing trends using statistical plots.

## 3.2 Summary Statistics

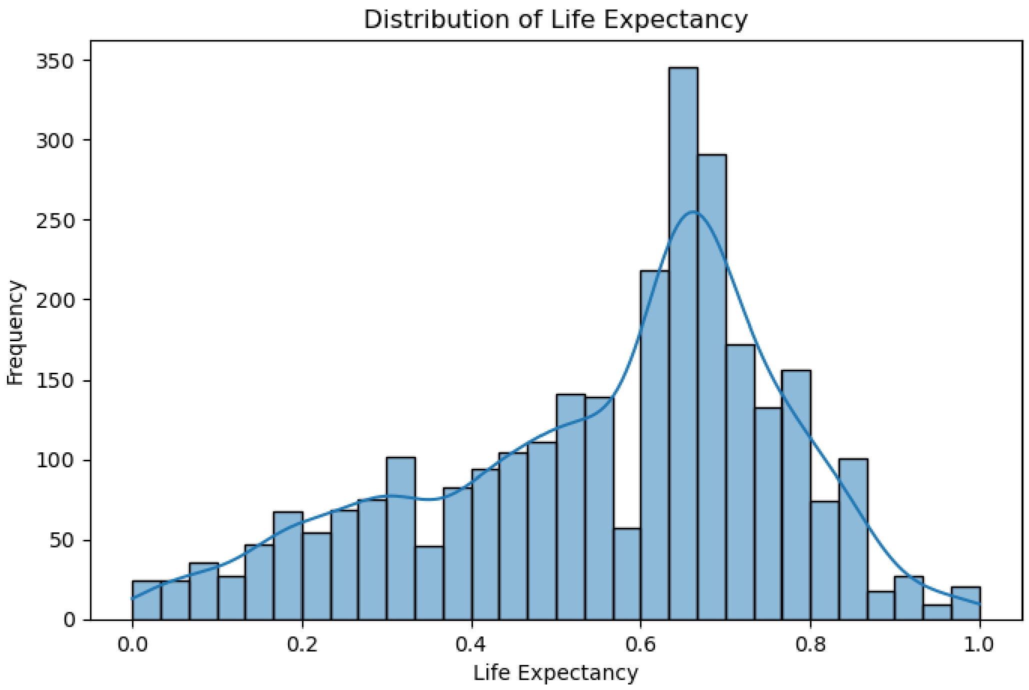
We began by generating **summary statistics** of the numerical features using `.describe()` to understand the dataset's distribution.

**Key Insights:**

- **Life Expectancy** ranged from **36 years to 89 years**.
- **GDP values** varied significantly, confirming the need for standardization.
- **Infant mortality rates** were notably higher in developing countries.

## 3.3 Distribution of Target Variable (Life Expectancy)

To check the **distribution of life expectancy**, we used a **histogram and KDE (Kernel Density Estimation) plot**.

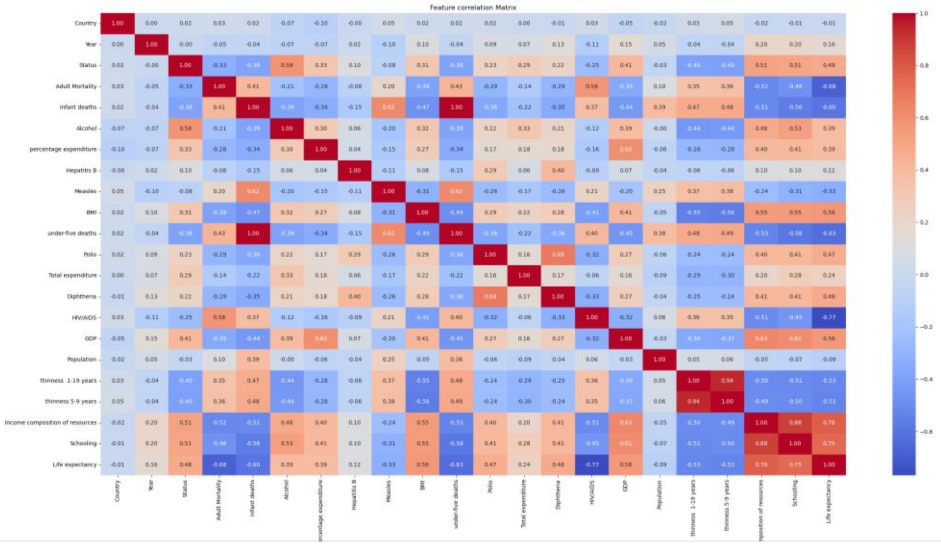


**Observations:**

- The **distribution is slightly right-skewed**, meaning some countries have significantly higher life expectancy.
- The majority of life expectancy values lie between **60 to 80 years**.

## 3.4 Correlation Analysis

To identify relationships between **Life Expectancy** and other features, we calculated the Pearson correlation coefficient.



**Key Findings:**

- **Life Expectancy has a strong positive correlation** with:
  - **Income Composition of Resources (0.78)**
  - **Schooling (0.74)**
  - **BMI (0.56)**
  - **GDP (0.55)**
- **Life Expectancy is negatively correlated** with:
  - **HIV/AIDS (-0.76)**
  - **Adult Mortality (-0.68)**
  - **Infant Deaths (-0.59)**
  - **Under-five Deaths (-0.62)**

**Implication:** Countries with **higher GDP, better schooling, and better nutrition (BMI)** tend to have **higher life expectancy**, while **high mortality rates and diseases (HIV/AIDS, Infant Deaths)** decrease life expectancy.

## 3.5 Boxplots to Detect Outliers

To detect outliers in key features, we used boxplots.

**Observations:**

- **GDP has extreme outliers**, likely due to developed countries with very high GDP.
- **Infant Deaths and HIV/AIDS show high skewness**, confirming the need for transformations.

## 3.6 Comparing Life Expectancy for Developed vs. Developing Countries

We visualized the difference in life expectancy between **Developed** and **Developing** countries.

- **Developed countries** have significantly higher life expectancy.

- **Developing countries** show a wider range of life expectancy values.

## 3.7 Relationship Between GDP and Life Expectancy

To analyze the impact of economic factors, we plotted **GDP vs. Life Expectancy**.

**Observations:**

- A **positive trend** is visible: **higher GDP → higher Life Expectancy**.
- Some outliers exist where high GDP does not necessarily mean high life expectancy (e.g., oil-rich nations).

## 3.8 Conclusion from EDA

From our exploratory analysis, we derived the following key insights:

- **Countries with higher GDP, better schooling, and better nutrition tend to have higher life expectancy.**
- **Developing countries exhibit greater variability in life expectancy.**
- **HIV/AIDS, infant mortality, and poor healthcare access negatively impact life expectancy.**
- **Economic growth (GDP) positively correlates with longer life expectancy.**

# 4. Statistical Analysis

Statistical analysis helps in validating our observations from EDA using hypothesis testing.

## 4.1 Descriptive Statistics

Descriptive statistics summarize key properties of the dataset. We use `.describe()` to compute measures like mean, median, variance, and standard deviation.

## 4.2 Inferentail Statistics

### 1. Confidence Interval for Life Expectancy (95% CI: (68.98, 69.66))

- A **confidence interval** provides a range in which the true population mean is expected to lie with 95% confidence.
- This means that, based on your sample, the **true population mean life expectancy is likely between 68.98 and 69.66 years**.

## 2. One-Sample T-Test (Comparing Sample Mean with Population Mean)

- **Sample Mean = 67.33, Population Mean = 69.32**
- **T-Statistic = -1.0165, P-Value = 0.3178**
- **Fail to Reject Null Hypothesis** → No significant difference between sample and population mean.

**Key Points:**

- A **t-test** checks if the sample mean differs significantly from the population mean.
- A **high p-value (0.3178 > 0.05)** means the difference is not statistically significant.
- This suggests that the sample life expectancy is **not significantly different from the population life expectancy**.

## 3. Chi-Square Test (Association Between Life Expectancy and Country Status)

- **Chi-Square Statistic = 503.9067, P-Value = 0.00000**
- **Reject Null Hypothesis** → Life expectancy **is significantly associated with country status (Developed/Developing)**.

**Key Points:**

- A **Chi-square test** checks if two categorical variables are independent.
- A **p-value of 0.00000 (< 0.05)** means the association is statistically significant.
- This indicates that **country status (Developed vs. Developing) strongly influences life expectancy**.

## 4. ANOVA Test (Effect of GDP on Life Expectancy)

- **ANOVA Statistic = 576.67, P-Value = 3.50e-211**
- **Reject Null Hypothesis** → **GDP significantly affects life expectancy**.

**Key Points:**

- ANOVA tests **whether the means of multiple groups differ significantly**.
- A **very low p-value (almost zero)** suggests a strong impact.

- This means **GDP has a significant effect on life expectancy**, likely indicating that **higher GDP leads to higher life expectancy**.

## 4.3 Key Conclusions from Statistical Analysis

- **Confidence Interval (95%)**: The true population mean life expectancy is likely between **68.98 and 69.66 years**.
- **T-Test**: No significant difference between **sample mean (67.33)** and **population mean (69.32)** (*p=0.3178p = 0.3178*p=0.3178, fail to reject *H0H_0*H0 ).
- **Chi-Square Test**: Life expectancy is **significantly associated** with a country's development status (*p<0.00001p < 0.00001*p<0.00001, reject *H0H_0*H0 ).
- **ANOVA Test**: GDP has a **significant impact** on life expectancy (*p≈0p \approx 0*p≈0, reject *H0H_0*H0 ).

## Conclusion & Final Takeaways

- **Economic growth, education, and healthcare access** are the strongest **drivers of life expectancy**.
- **Preventable diseases, malnutrition, and mortality rates must be addressed** to improve life expectancy in developing countries.
- **Government policies focusing on healthcare infrastructure, vaccination programs, and nutrition** can drastically improve life expectancy worldwide.

Sources of Dataset :- https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Sources used for handling Outliers :-

https://youtu.be/4FBPtgLbehY?si=m8ul_65lfeafApgo

https://youtu.be/jmAuVP_UOn0?si=F3mR-d3t0gH-_iBU