**WEEK 1 GROUP ASSIGNMENT - EXPLORATORY DATA ANALYSIS AND DATA CLEANING**

GROUP MEMBERS – SWARNA DHARSHINI S, RAJINDEEP PHARWAHA, PRIYANKA S

Dataset Link -[**life-expectancy-and-socio-economic-world-bank**](https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank)

1. **PROBLEM STATEMENT**
   1. **Problem Statement and Dataset**

The problem we have chosen to address is understanding the factors that influence life expectancy across different countries globally. The dataset we are utilizing is titled "Life Expectancy & Socio-Economic" from the World Bank, sourced from Kaggle. It contains data spanning from 2000 to 2019, providing insight into various socio-economic factors and their potential impact on life expectancy.

* 1. **Why this problem is important to solve?**

Addressing the issue of life expectancy linked to socio-economic factors is crucial for creating a more equitable, prosperous, and stable world. Understanding these factors can help improve global health outcomes while tackling larger challenges such as poverty, inequality, and global security. Solving this problem could lead to better health interventions, which can enhance quality of life and contribute to broader societal benefits.

* 1. **Who would benefit from the solution?**

The solution would benefit a broad range of stakeholders, from local populations to international bodies. Poorer countries could use these insights to improve access to basic necessities like clean water, nutritious food, healthcare, and education. This would not only raise living standards but also reduce the need for migration in search of better conditions. Additionally, it would aid public health services worldwide by mitigating disease outbreaks and improving overall global health.

* 1. **Who are our audience or who your stakeholders might be?**

The primary audience for this analysis would be governments around the world. By understanding the socio-economic factors affecting life expectancy, governments can make informed decisions about investing in services that improve health outcomes. These insights can guide public policy to address the root causes of low life expectancy and drive sustainable improvements in national health and well-being.

1. **DATASET**
   1. **Chosen Dataset and reason for choosing it**

We have chosen the "Life Expectancy & Socio-Economic" dataset from the World Bank, available on Kaggle. This dataset provides comprehensive global data on life expectancy, along with various socio-economic indicators that influence it. Covering the period from 2000 to 2019, the dataset includes information on critical factors such as health expenditure, education expenditure, undernourishment prevalence, and unemployment rates. We selected this dataset because it offers rich, detailed information suitable for analysing the relationship between socio-economic conditions and life expectancy. Additionally, it contains missing values, allowing us to apply data cleaning techniques to prepare it for deeper analysis.

* 1. **Dataset Overview**

This dataset provides the necessary features to examine how socio-economic factors influence life expectancy, offering both the complexity and richness required for our analysis.

* **Country Name** (object): The name of the country.
* **Country Code** (object): A unique code for each country.
* **Region** (object): The geographical region to which the country belongs.
* **Income Group** (object): The income classification of the country.
* **Year** (int64): The year of the data point.
* **Life Expectancy (World Bank)** (float64): Life expectancy at birth in years.
* **Prevalence of Undernourishment** (float64): The percentage of the population suffering from undernourishment.
* **CO2** (float64): Carbon dioxide emissions in metric tons per capita.
* **Health Expenditure %** (float64): Health expenditure as a percentage of GDP.
* **Education Expenditure %** (float64): Education expenditure as a percentage of GDP.
* **Unemployment** (float64): The unemployment rate as a percentage of the total labor force.
* **Corruption** (float64): A measure of corruption within the country.
* **Sanitation** (float64): Access to improved sanitation facilities, expressed as a percentage of the population.
* **Injuries** (float64): Number of deaths due to injuries.
* **Communicable** (float64): Deaths due to communicable diseases.
* **Non-Communicable** (float64): Deaths due to non-communicable diseases.
  1. **How Does It Fit the Problem?**

This dataset is particularly suitable for addressing our problem as it offers a rich set of socio-economic indicators that directly or indirectly affect life expectancy. By analysing factors like health expenditure, access to education, and undernourishment, we can investigate their relationship with life expectancy and how they vary across different regions and income groups. The dataset also contains global coverage, making it ideal for comparing diverse regions and identifying trends or disparities in life expectancy due to socio-economic differences. Additionally, the presence of missing values allows us to apply data cleaning techniques, an essential part of the data analytics process.

1. **DATA CLEANING PROCESS**
   1. **Handling Missing Data:**
2. **Remove Columns with Too Many Missing Values**

data.dropna(axis=1, thresh=len(data) 0.5, inplace=True)

**Reason for Use:**

- Columns with more than 50% missing values are removed because they provide insufficient information for meaningful analysis and may introduce bias.

- Columns with excessively high missing values are likely to be unreliable and contribute little to the analysis. Removing these columns allows us to focus on features with more complete data, enhancing the overall quality and reliability of our dataset.

1. **Mean Imputation for 'Health Expenditure %' and 'Education Expenditure %' Columns**

data['Health Expenditure %'].fillna(data['Health Expenditure %'].mean(), inplace=True)

data['EducationExpenditure %'].fillna(data['Education Expenditure %'].mean(), inplace=True)

**Reason for Use:**

- Health Expenditure %: This continuous variable is expected to have a relatively symmetric distribution. Mean imputation is appropriate as it assumes the missing data is missing at random (MAR) and provides a central value that is straightforward to calculate.

- Education Expenditure %: Similar to Health Expenditure, this variable benefits from mean imputation if its distribution is not heavily skewed. Mean imputation offers a simple and effective method for handling missing values by substituting them with the mean of the observed data.

1. **Median Imputation for 'Unemployment' Column**

data['Unemployment'].fillna(data['Unemployment'].median(), inplace=True)

**Reason for Use:**

- Unemployment: This column may exhibit a skewed distribution or contain outliers. Median imputation is more robust in such cases because it is less influenced by extreme values, providing a more stable estimate of central tendency for missing values.

1. **Mode Imputation for 'Prevalence of Undernourishment' Column**

data["PrevalanceofUndernourishment"].fillna(data["PrevalanceofUndernourishment"] .mode()[0], inplace=True)

**Reason for Use:**

- Prevalence of Undernourishment: This column represents discrete numerical values where the most frequent value (mode) serves as a good estimate for missing data. Mode imputation is suitable when missing values are assumed to follow the most common observed pattern, making it an appropriate method for categorical or discrete data.

1. **Group-Based Imputation for 'Sanitation' Column**

data['Sanitation']=data.groupby('Region')['Sanitation'].transform(lambdax:x.fillna(x.mean()))

**Reason for Use:**

- Sanitation: Missing values in this column are imputed based on the average sanitation values within the same region. This method is justified because different regions may have distinct sanitation levels, and imputing based on regional averages provides a more contextually accurate estimate.

1. **Interpolation for 'Life Expectancy World Bank' and 'CO2' Columns**

data['Life Expectancy World Bank'] = data['Life Expectancy World Bank'].interpolate()

data['CO2'] = data['CO2'].interpolate()

**Reason for Use:**

- Life Expectancy World Bank: This column, which likely represents time-series or sequential data, benefits from interpolation as it estimates missing values based on trends and continuity in the data. Interpolation provides a smooth estimate that reflects the gradual changes in life expectancy over time.

- CO2: Similar to Life Expectancy, CO2 levels may change gradually. Interpolation helps in maintaining continuity and accurately estimating missing values based on the surrounding data points.

* 1. **Dealing with Outliers:**
     1. **Identifying Outliers:**
* **Z-Score Method:** This method measures how far a data point is from the mean in terms of standard deviations. A z-score greater than 3 or less than -3 is often considered an outlier.
* **IQR (Interquartile Range) Method:** Outliers are values that lie below the first quartile (Q1) minus 1.5 times the interquartile range (IQR) or above the third quartile (Q3) plus 1.5 times the IQR.
  + 1. **Handling Outliers:**
  + **Capping (Winsorization):** Extreme values are capped at a certain percentile, often 1st and 99th, to reduce their influence while preserving the rest of the data.

data['CO2']=data['CO2'].clip(lower=data['CO2'].quantile(0.01),upper=data['CO2']. quantile(0.99))

data['Health Expenditure %'] = data['Health Expenditure %'].clip(lower=data['Health Expenditure %'].quantile(0.01), upper=data['Health Expenditure %'].quantile(0.99))

* 1. **Data Consistency:**

Data consistency involves ensuring that the data is uniform, follows the same units, and conforms to expected formats. Inconsistent data can lead to incorrect conclusions and unreliable models.

**Methods for Ensuring Consistency:**

* **Data Type Checks:** Verify that each column has the correct data type, and convert data types where necessary.

data['Year'] = data['Year'].astype(int)

* **Handling Inconsistent Formats:** For example, some columns may have percentage values stored as decimals and others as whole numbers. Standardize these formats.

data['Health Expenditure %'] = data['Health Expenditure %'].apply(lambda x: x\*100 if x < 1 else x)

* **Ensuring Uniform Units:** Make sure that columns representing the same type of data use uniform units (e.g., CO2 emissions measured in the same unit across all records).

data['CO2'] = data['CO2'].apply(lambda x: x \* 1000 if x < 1000 else x) # Example if CO2 is given in different units

* **Checking for Duplicates:** Remove any duplicated records to ensure data is not being analyzed multiple times.

data.drop\_duplicates(inplace=True)

1. **EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis (EDA) involves understanding the structure of the data, identifying key features, discovering patterns, and detecting any potential challenges. In this section, we explore different aspects of the dataset to uncover insights into how various socio-economic factors relate to life expectancy.

* 1. **Regional Contributions Without Corruption**

To further examine the breakdown of contributing factors excluding **Corruption**, we generated a stacked bar chart for the same, but with corruption removed from the list of factors. This analysis provides insights into other factors affecting life expectancy in the absence of corruption data.

# New factors list without corruption

factors2 = ['Unemployment', 'Prevelance of Undernourishment', 'CO2', 'Health Expenditure %', 'Education Expenditure %', 'Sanitation', 'Injuries', 'Communicable', 'NonCommunicable']

# Group data by Region and count the total for each contributing factor

grouped\_data2 = data.groupby('Region')[factors2].count()

# Plot a stacked bar chart

grouped\_data2.plot(kind='bar', stacked=True, figsize=(8, 8), colormap='Dark2')

# Label axes

plt.xlabel('Region')

plt.ylabel('Total Number of Contributing Factors')

# Add grid, legend, and layout adjustment

plt.title('Breakdown of Contributing Factors by Region')

plt.legend(fontsize="small", title="Factors")

plt.grid(linestyle='--')

plt.tight\_layout()

plt.show()

* 1. **Distribution of Life Expectancy**

A histogram of **Life Expectancy (World Bank)** was plotted to understand the distribution of life expectancy values in the dataset. This helps in identifying the range, frequency, and any skewness in life expectancy across the different countries.

# Histogram of Life Expectancy by Frequency in the dataset

data['Life Expectancy World Bank'].plot(kind='hist', bins=20, title='Life Expectancy World Bank')

plt.show()

This histogram allows us to observe whether the life expectancy values are normally distributed or skewed and if there are any clear outliers.

* 1. **Available Records by Region**

The following bar chart visualizes the total number of available records per region. This gives an overview of how much data is available for each region, highlighting regions that may have insufficient data for comprehensive analysis.

# Plot showing the total number of records available for each region

data.groupby('Region').size().plot(kind='barh', color=sns.palettes.mpl\_palette('Dark2'))

plt.title('Available Records by Region')

plt.grid(linestyle='--')

plt.tight\_layout()

plt.show()

* 1. **Relationship Between Socio-Economic Factors and Life Expectancy**

We generated scatter plots to explore the relationship between various socio-economic factors and **Life Expectancy (World Bank)**. Each plot is color-coded by region to observe potential patterns or trends by geographic area. These scatter plots help us understand how different factors, such as unemployment or healthcare expenditure, influence life expectancy.

# List of factors

factors = ['Unemployment', 'Prevelance of Undernourishment', 'CO2', 'Health Expenditure %', 'Education Expenditure %', 'Sanitation', 'Injuries', 'Communicable', 'NonCommunicable']

# Loop through each factor and create scatter plots against Life Expectancy World Bank

for factor in factors:

plt.figure(figsize=(8, 6))

sns.scatterplot(data=data, x=factor, y='Life Expectancy World Bank', hue='Region', palette='Set1')

# Add labels and title

plt.xlabel(factor)

plt.ylabel('Life Expectancy World Bank')

plt.title(f'{factor} vs. Life Expectancy')

plt.grid(True)

plt.show()

* 1. **Breakdown of Contributing Factors by Region**

We grouped the data by **Region** to examine the number of records for each contributing factor across different regions. This gives an overview of the distribution of data and highlights which regions are better represented in the dataset. The stacked bar charts allow us to compare the total number of records for each factor in various regions.

# Group data by Region and count the total for each contributing factor

grouped\_data = data.groupby('Region')[factors].count()

# Plot a stacked bar chart

grouped\_data.plot(kind='bar', stacked=True, figsize=(8, 8), colormap='Dark2')

# Label axes

plt.xlabel('Region')

plt.ylabel('Total Number of Contributing Factors')

# Add grid, legend, and layout adjustment

plt.title('Breakdown of Contributing Factors by Region')

plt.legend(fontsize="small", title="Factors")

plt.grid(linestyle='--')

plt.tight\_layout()

plt.show()

This chart helps in understanding the distribution of data and whether any region lacks data for key contributing factors. It is also useful in identifying regions with extensive data, which might be useful for deeper analysis.

* 1. **Data Quality Overview**

In addition to analyzing the relationships between variables, we also assessed the quality of the data. This includes:

* **Missing Data:** Missing values were addressed using various imputation techniques, ensuring data completeness for the analysis.
* **Outliers:** Outliers were detected and handled appropriately to prevent them from distorting the results.
* **Data Consistency:** Efforts were made to maintain consistency across different regions and time periods, ensuring a uniform approach to how socio-economic factors are measured.