**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**

[Task 1: Identify a real-world problem you are interested in solving using data analytics. You  
should justify why this problem is important to solve, and who would benefit from the  
solution. In other words, think about your audience or who your stakeholders might be. It is  
up to you what kind of problem you want to choose. It can be something you are interested,  
something fun, something that relates to the industry you are most interested in, or all of  
them.  
Note: Before researching datasets, you should think about potential issues that you can  
encounter. For example, if you work on a healthcare problem, you might not be able to find  
openly available patient level data.]

* 1. **Problem Statement and Dataset**

The problem we have chosen to address is what are the factors affecting life expectancy in different countries globally. The dataset we are using is **Life expectancy & Socio-Economic (world bank)** which we found on **Kaggle**. The data runs from 2000 – 2019.

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

Solving the problem of life expectancy linked to socio-economic factors is very important as it creates a more just, prosperous and stable world. It could improve health outcomes as well as addressing broader challenges, such as poverty, inequality and global security.

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

Solving this problem would benefit everyone across the world. It would assist poorer countries by giving their citizens better access to clean water, healthier foods, accessible healthcare and education. This would reduce migration from these countries to more affluent countries by people seeking better healthcare and living conditions. It would also benefit public health services by reducing disease outbreaks which could affect global health.

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

Our audience for this would be governments across the world to help them make informed decisions about which services to invest in which could improve their socio-economic factors that contribute to lower life expectancy.

1. **DATASET**

[Task 2: Based on the problem identified, research and find a dataset that can help address  
this problem. The dataset should be rich enough for complex analysis, including various  
types of data and, preferably, some missing or dirty data. You might find yourself in the  
situation where only one dataset might not be enough and that you may need to combine  
multiple datasets together.  
Note: You should explain the data source and the reasons you chose this particular  
dataset(s). How does it fit your problem?]

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

The dataset we chose to use was **Life expectancy & Socio-Economic (world bank)** which we found on **Kaggle.** This dataset has global data on life expectancy and which socio-economic factors affect life expectancy in different areas. The data runs from 2000 until 2019. As we were looking to compare different regions with the factors which affect life expectancy, this dataset appeared to contain a large amount of information which was relevant to our needs. There was also a lot of missing values which would require analysing and cleaning.

* 1. **AAA**
  2. **AAA**

1. **DATA CLEANING PROCESS**

[Task 4: Perform and document the data cleaning process, explaining why you made certain  
decisions. This should involve handling missing data, dealing with outliers, and ensuring  
data consistency.]

* 1. **Handling Missing Data:**

1. **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:**

When analysing the data, we found that there weren’t really any outliers which would affect the data. This is shown in the pairplot which shows all the different factors and how they compare with Life Expectancy based on region.

* 1. **Data Consistency:**

1. **EXPLORATORY DATA ANALYSIS**

[Task 3: Conduct an in-depth Exploratory Data Analysis (EDA) on your dataset(s), identify  
key features, any interesting patterns or anomalies, and potential challenges in the data  
(missing values, outliers, etc.). Note: You should also give an overview over the data quality.]

* 1. **AAAA**
  2. **AAAA**
  3. **AAAA**