**WEEK 3 GROUP ASSIGNMENT - Data Visualisation and Communication**

GROUP MEMBERS – Swarna Dharshini S, Rajindeep Pharwaha, Marina Klepaci, Priyanka S

**Problem Formulation:**

The problem we chose to address was what are the factors affecting life expectancy in different countries globally. Addressing this issue is crucial for creating a more equitable, prosperous, and stable world. It could improve health outcomes as well as addressing broader challenges, such as poverty, inequality and global security.

Solving this problem would benefit everybody globally. 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.

This solution would benefit a wide range of stakeholders, from local populations to international organisations. Poorer countries could use these insights to improve access to essentials like clean water, food, healthcare, and education, raising living standards and reducing migration. It would also help public health services globally by preventing disease outbreaks and enhancing overall health.

**Data Collection:**

We chose the **Life Expectancy & Socio-Economic(world bank)** dataset from Kaggle, which contains 3,306 records from 174 countries between 2000 and 2019, on global life expectancy and its socio-economic factors. We decided to use a Google Colab Notebook to carry out Data Analysis and Cleaning.

**Data Quality:**

The quality of the 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.

There were missing values within the Corruption(**71%**), Sanitation(**38%**), Education Expenditure(**33%**), Prevalence of Undernourishment(**21%**), Life Expectancy(**6%**) and Health Expenditure(**5%**) columns.

**Analysis:**

During the Analysis process, we used a Google Colab notebook to analyse and clean the dataset. This was done using a number of techniques including dropping columns which had too many missing values as they would provide insufficient information for meaningful analysis and may introduce bias. Due to the large number of missing values in the corruption column, it was decided to drop this column completely as it would not give any meaningful analysis. **NaN** valueswere also replaced by using either **mean, median or mode** calculated values depending on the type of data within the column. Outliers were capped to reduce their influence while preserving the rest of the data. A number of graphs were also produced to analyse the dataset including a **pair plot** to show correlations between global life expectancy and each of the factors affecting it, as well as a **histogram** to show the frequency of global life expectancy ages in the dataset, and a **bar graph** showing the total number of records held for each region.

A new cleaned excel sheet was then produced to carry out further statistical analysis including a more detailed look at which factors affect global life expectancy the most. From this analysis we could see that the **prevalence of undernourishment** and **unemployment** negatively affected life expectancy the most while high levels of **sanitation** positively impacted life expectancy. A **correlation matrix** was produced to show how each factor affected life expectancy as well as graphs to show **life expectancy** against **prevalence of undernourishment** as well as **life expectancy** against **unemployment**.

We created a dashboard to present our findings, beginning with a **choropleth** **map** that illustrates the global distribution of **life** **expectancy**. A **histogram** follows, displaying the frequency of different **life** **expectancy** ranges in the dataset. Next, a **correlation** **heatmap** highlights the relationships between socio-economic factors and **life** **expectancy**, revealing that **prevalence of undernourishment** has the strongest **negative** **impact**, while good **sanitation** has the strongest positive effect. The **hexbin plot** and **bar graph** further explore the link between **sanitation** and **life expectancy**, clearly showing that better **sanitation** correlates with **longer life expectancy**. A **scatter plot** then shows the regional correlation between **prevalence of** **undernourishment** and **life** **expectancy**. The dashboard concludes with a **box** **plot** and time series plot, illustrating the distribution of **life** **expectancy** by region and its changes over time.

From our analysis we can see that about **29%** of the variation in life expectancy can be explained by the prevalence of undernourishment. In contrast, there is a **strong, statistically significant positive relationship** between sanitation and life expectancy. **About 45% of the variation in life expectancy** can be explained by sanitation levels alone, showing that sanitation is a crucial factor in improving life expectancy. There is a **positive but weak relationship** between health expenditure as a percentage of GDP and life expectancy. A 1% increase in health expenditure is associated with an increase of about **6.36 years** in life expectancy, but the **R-squared is quite low**, explaining only about **8.32%** of the variability in life expectancy. There is a **statistically significant but weak positive relationship** between education expenditure and life expectancy. A 1% increase in education expenditure is associated with an increase of **4.59 years** in life expectancy.

There have been many studies conducted on global life expectancy the factors affecting it, however, these only focus on one specific area. For example, [Life Expectancy - Our World in Data](https://ourworldindata.org/life-expectancy?insight=there-are-wide-differences-in-life-expectancy-around-the-world#key-insights) explores how life expectancy has increased globally over the years. Whereas, [Impacts from Economic Development and Environmental Factors on Life Expectancy: A Comparative Study Based on Data from Both Developed and Developing Countries from 2004 to 2016 - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8391297/) is a study of 20 developed and developing countries, exploring 9 economic and environmental indicators, including **CO2 levels**, **GDP**, **healthcare expenditure** and **education** **expenditure** between 2004 and 2016. Both studies demonstrate that while global life expectancy has improved over the years, several factors remain that, if addressed, could significantly boost life expectancy in the future.

**Recommendations:**

There are several recommendations which can be made to improve life expectancy globally. These are listed below: -

* **Enhance Sanitation Infrastructure:**
  + **Action:** Prioritize investments in sanitation facilities, particularly in developing regions. Initiatives should include building clean water supply systems and effective sewage management.
* **Expected Outcome:** Improved sanitation is likely to lead to significant increases in life expectancy by reducing waterborne diseases. Countries with over 70% access to clean water report life expectancies of 77 years or more, while those below 60% access only average 69 years. For every **unit improvement in sanitation**, life expectancy increases by approximately **36.29 years**.
* **Address Undernourishment:**
  + **Action:** Implement targeted nutrition programs for vulnerable populations, such as school meal initiatives and community health education on nutrition.
* **Expected Outcome:** Reducing malnutrition will directly enhance health outcomes and contribute to increased life expectancy. **Every 1% increase in undernourishment** leads to a **0.47 year decrease** in life expectancy.
* **Increase Health Expenditure:**
  + **Action:**  Allocate additional resources to healthcare services, focusing on preventive care and emergency services. This includes enhancing the capacity of health facilities and increasing public health campaign budgets.
* **Expected Outcome:** Higher health expenditure is expected to improve health metrics and enhance life expectancy.
* **Explore Education Initiatives:**
  + **Action:** Invest in health literacy programs, especially in areas with low educational attainment. Long-term strategies should aim to improve access to quality education for all demographics.
* **Expected Outcome:** While the immediate effects on life expectancy may be limited, better education can promote healthier lifestyles and informed health decisions, leading to improved health outcomes over time.