**WEEK 2 GROUP ASSIGNMENT - Explanatory Data Analysis and Statistical Inference**

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

1. **UNIVARIATE AND BIVARIATE ANALYSES**

The goal of this analysis is to explore the dataset on life expectancy and socio-economic factors to uncover trends and relationships that may inform public health policies. By examining the distribution of life expectancy and its relationship with various socio-economic variables, we can gain insights into factors that significantly impact health outcomes across different countries.

* 1. **Univariate Analysis: Life Expectancy Histogram**
* The histogram of life expectancy reveals the distribution of life expectancy values across countries.
* A significant portion of countries falls within the range of **60 to 80 years**, indicating that most countries have a life expectancy that aligns with global averages.
* There are noticeable peaks in the histogram, suggesting clusters of countries with similar life expectancies, possibly influenced by socio-economic factors such as healthcare access and education levels.
* The histogram may also indicate a right-skewed distribution, where fewer countries experience very high life expectancies, emphasizing disparities in health outcomes globally.
  1. **Bivariate Analysis: Scatter Plots**
* The scatter plot comparing **Life Expectancy and Health Expenditure %** shows a positive trend, suggesting that as health expenditure increases, life expectancy tends to increase as well.
* This relationship indicates the importance of healthcare investment in improving population health. Countries with higher health expenditure often report better health outcomes and longer life spans.
* However, the scatter plot may also reveal variability, with some countries achieving high life expectancy despite lower health expenditure, potentially indicating other influencing factors like sanitation and nutrition.
* Outliers may exist, where certain countries demonstrate high life expectancy with low health expenditure, highlighting successful public health interventions or social policies that merit further investigation.
  1. **Grouped Analysis: Stacked Bar Chart by Region**
* The stacked bar chart illustrates the breakdown of life expectancy and unemployment by region, providing a comparative view of how these factors vary across different geographic areas.
* Regions with high life expectancy often correspond to lower unemployment rates, suggesting a correlation between economic stability and health outcomes.
  + 1. **Regional Disparities in Life Expectancy:**
* The analysis reveals that **Europe & Central Asia** has the highest total life expectancy at approximately **67,694.07 years**, indicating a robust healthcare system and favourable socio-economic conditions.
* In contrast, **North America**, despite being economically developed, shows a relatively low total life expectancy of **4,542.13 years**, highlighting the need for further investigation into the calculation or specific conditions affecting this region.
  + 1. **Correlation Between Life Expectancy and Unemployment:**
* Regions with higher total life expectancy generally have lower unemployment figures. For example, **East Asia & Pacific** and **Europe & Central Asia** exhibit lower unemployment rates (2,190.46 and 8,531.73, respectively), correlating with their higher life expectancy.
* **Sub-Saharan Africa** shows a concerning trend with **48,105.67 years** of life expectancy coupled with a higher unemployment rate of **6,915.16**. This suggests socio-economic challenges that may hinder health outcomes and overall quality of life.
  + 1. **Latin America & Caribbean vs. South Asia:**
* Both regions have comparable life expectancy totals, yet **Latin America & Caribbean** has a higher total of **40,487.87 years** against a unemployment figure of **3,933.67** compared to **South Asia’s** **10,416.19 years** and **773.44** unemployment. This indicates potential disparities in health services, education, and economic stability within South Asia.

1. **HYPOTHESIS TESTING IN LIFE EXPECTANCY ANALYSIS**

The following hypotheses have been identified:

* 1. **Life Expectancy vs. Health Expenditure %**

Hypothesis 1: Higher health expenditure as a percentage of GDP is associated with higher life expectancy.

* Regression Statistics:
  + R² = 0.0832: Approximately 8.32% of the variance in life expectancy can be explained by health expenditure.
  + P-value < 0.0001: The result is statistically significant.
* Interpretation: Higher health expenditure correlates with increased life expectancy, confirming Hypothesis 1.
  1. **Life Expectancy vs. Prevalence of Undernourishment**

Hypothesis 2: There is a negative correlation between the prevalence of undernourishment and life expectancy.

* Regression Statistics:
  + R² = 0.2925: About 29.25% of the variance in life expectancy is explained by the prevalence of undernourishment.
  + P-value < 0.0001: Highly significant result.
* Interpretation: A higher prevalence of undernourishment is associated with lower life expectancy, supporting Hypothesis 2.
  1. **Life Expectancy vs. Education Expenditure %**

Hypothesis 3: Increased education expenditure is positively correlated with life expectancy.

* Regression Statistics:
  + R² = 0.0236: Only 2.36% of the variance in life expectancy can be explained by education expenditure.
  + P-value < 0.0001: Statistically significant.
* Interpretation: While there is a positive correlation, the impact of education expenditure on life expectancy is relatively weak, indicating partial support for Hypothesis 3.
  1. **Life Expectancy vs. Sanitation**

Hypothesis 4: Higher sanitation levels correlate positively with life expectancy.

* Regression Statistics:
  + R² = 0.4505: 45.05% of the variance in life expectancy can be explained by sanitation.
  + P-value < 0.0001: Extremely significant result.
* Interpretation: Improved sanitation strongly correlates with increased life expectancy, validating Hypothesis 4.
  1. **Life Expectancy vs. Unemployment**

Hypothesis 5: Unemployment rates have a negative impact on life expectancy.

* Regression Statistics:
  + R² = 0.0035: Very low explanatory power.
  + P-value = 0.0007: Statistically significant.
* Interpretation: Although unemployment has a statistically significant negative relationship with life expectancy, the weak correlation suggests its limited impact, partially supporting Hypothesis 5.

1. **CORRELATION ANALYSIS**
   1. **Life Expectancy and Prevalence of Undernourishment (-0.541)**

* There is a **moderately strong negative correlation** between life expectancy and the prevalence of undernourishment. This means that as undernourishment increases, life expectancy tends to decrease. This aligns with expectations as malnutrition impacts health and survival, especially in developing regions.
  1. **Life Expectancy and CO2 Emissions (0.038)**
* The correlation between life expectancy and CO2 emissions is **very weak (close to zero)** and positive. This suggests almost no linear relationship between these two variables. CO2 emissions are more likely associated with industrialization or economic factors, but their direct effect on life expectancy might be minimal or indirectly related.
  1. **Life Expectancy and Health Expenditure % (0.288)**
* There is a **weak positive correlation** between life expectancy and the percentage of health expenditure. While the correlation is not very strong, it does suggest that countries spending a higher percentage of their GDP on health tend to have higher life expectancies.
  1. **Life Expectancy and Education Expenditure % (0.154)**
* The correlation between life expectancy and education expenditure is also **weakly positive**. This may suggest that countries investing more in education may indirectly promote longer life expectancy, possibly through increased awareness and healthier lifestyles, although the correlation is weak.
  1. **Life Expectancy and Unemployment (-0.059)**
* The correlation between life expectancy and unemployment is **very weak and negative**. While unemployment is generally associated with economic distress and poorer health outcomes, this weak correlation indicates a minimal direct effect in this dataset.
  1. **Life Expectancy and Sanitation (0.671)**
* There is a **strong positive correlation** between life expectancy and access to sanitation. This is one of the strongest relationships in the table, which is expected because improved sanitation reduces the spread of infectious diseases and improves public health, directly impacting life expectancy.
  1. **Life Expectancy and Injuries (-0.006)**
* The correlation between life expectancy and injuries is essentially **zero**, suggesting that the overall impact of injuries on life expectancy is not significant in this dataset.
  1. **Life Expectancy and Communicable Diseases (-0.215)**
* There is a **moderate negative correlation** between life expectancy and communicable diseases. As the prevalence of communicable diseases increases, life expectancy tends to decrease. This is in line with global health trends, as higher rates of communicable diseases often reduce life expectancy, especially in low-income countries.
  1. **Life Expectancy and Non-Communicable Diseases (0.044)**
* There is a **very weak positive correlation** between life expectancy and non-communicable diseases. This is surprising, as non-communicable diseases are typically associated with higher mortality, especially in high-income countries. However, the weak positive correlation might suggest that better healthcare systems in countries with higher life expectancy might be managing these diseases more effectively.

1. **DATA INSIGHTS**
   1. **Insights and Key Takeaways**
      1. **Sanitation and Prevalence of Undernourishment:**

* **Correlation with Life Expectancy:**
  + Sanitation (Correlation: 0.671): A strong positive correlation indicates that improved sanitation is a crucial determinant of public health, particularly in developing countries. Effective sanitation reduces disease transmission, promoting better health outcomes.
  + Prevalence of Undernourishment (Correlation: -0.541): This significant negative correlation reflects the detrimental impact of malnutrition on health, highlighting its role as a major factor affecting life expectancy.
    1. **Health Expenditure:**
* **Correlation and Hypothesis Testing:** The hypothesis that higher health expenditure correlates with increased life expectancy was supported. Although the correlation is weaker than that for sanitation or nutrition, it underscores the necessity of investing in healthcare resources and access to improve public health outcomes.
  + 1. **Education Expenditure:**
* **Correlation and Hypothesis Testing:** While the hypothesis regarding the positive relationship between education expenditure and life expectancy was confirmed, the correlation was weak. This suggests that while educational investments may yield long-term health benefits, their immediate impact on life expectancy is limited.

CO2 Emissions, Unemployment, and Injuries:

* Correlation: These factors showed very weak correlations with life expectancy, indicating minimal direct impact. Their effects may be more indirect, highlighting the complexity of health determinants.
  1. **Recommendations**
     1. **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.
  + 1. **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.
  + 1. **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.
  + 1. **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.