**Life Expectancy: A Data-Driven Exploration and Prediction**

**Milestone: Data Collection and Pre-processing**

Group 13

Sai Varun Kumar Namburi

Prajwal Srinivas

[namburi.sai@northeastern.edu](mailto:namburi.sai@northeastern.edu)

[srinivas.pra@northeastern.edu](mailto:srinivas.pra@northeastern.edu)

**Percentage of Effort Contributed by Student1: 50%**

**Percentage of Effort Contributed by Student2: 50%**

**Signature of Student 1: Sai Varun**

**Signature of Student 2: Prajwal**

**Date of Submission: 12th Feb 2023**

**Data Source:** This dataset is taken from kaggle.com

<https://www.kaggle.com/code/philbowman212/life-expectancy-exploratory-data-analysis/data>

**Data Description:**

The World Health Organization's Global Health Observatory (GHO) maintains records of the health status and related factors of all countries. The data concerning life expectancy and health factors for 193 countries were obtained from the WHO's Global Health Observatory data repository. It was noted that over the past 15 years, there has been significant progress in the health sector, leading to a significant improvement in human mortality rates, particularly in developing nations compared to the last 30 years. In this project, data from the years 2000 to 2015 for 193 countries were selected for further analysis. In this dataset, we have 22 columns as below

(Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling)

**Data Collection:**

Life expectancy data has been gathered from a data repository called “Kaggle”. Information on 2938 records is captured in 22 columns. There is a total of 2 categorical and 20 numerical columns in the data.

**Columns Description:**

1. "Country": The name of the country is analyzed.
2. "Year": The year the data was collected.
3. "Status": Indicates the status of the country, such as "Developed" or "Developing".
4. "Life expectancy": The average number of years a person is expected to live in a specific geographic location.
5. "Adult Mortality": The number of deaths of individuals aged 15 to 60 years per 1,000 individuals in that age group.
6. "Infant Deaths": The number of deaths of infants under one year of age per 1,000 live births.
7. "Alcohol": The average amount of alcohol consumed by individuals in a specific geographic location.
8. "Percentage expenditure": The percentage of the government's total health expenditure spent on healthcare.
9. "Hepatitis B": The coverage rate of the hepatitis B vaccine in a specific geographic location.
10. "Measles": The number of reported cases of measles per 1,000 individuals.
11. "BMI": The average body mass index of individuals in a specific geographic location.
12. "Under-five Deaths": The number of deaths of children under five years of age per 1,000 live births.
13. "Polio": The coverage rate of the polio vaccine in a specific geographic location.
14. "Total expenditure": The total amount of money spent on healthcare by the government in a specific geographic location.
15. "Diphtheria": The coverage rate of the diphtheria vaccine in a specific geographic location.
16. "HIV/AIDS": The number of reported cases of HIV/AIDS in a specific geographic location.
17. "GDP": The gross domestic product of a specific geographic location.
18. "Population": The total population of a specific geographic location.
19. "Thinness 1-19 years": The percentage of individuals aged 1 to 19 years who are underweight in a specific geographic location.
20. "Thinness 5-9 years": The percentage of individuals aged 5 to 9 years who are underweight in a specific geographic location.
21. "Income composition of resources": The distribution of income in a specific geographic location.
22. "Schooling": The average number of years of schooling for individuals in a specific geographic location.

**Data Cleaning:**

Understanding the dataset:

Text

Description automatically generated

Check the null values present in each column by using the below command:

Text

Description automatically generated

Count of developing and developed countries from the dataset and compare the life expectancy and

Alcohol columns for the below countries.

A screenshot of a computer

Description automatically generated with medium confidence