Keylia Carter

IST 652

2/2/2023

Mini Project 1

World Health Organization (WHO): Life Expectancy

The World Health Organization aims to track external variables that would likely predict the health status of developing and developed countries. For this project, I am analyzing one target variable called Life\_Expectancy to better understand which of the 21 variables have a strong or weak relationship with the target variable. I’m also going to explore the descriptive statistics of different variables and illustrate my findings through data visualizations.

Below is a list of the variables found in the dataset:

1. **Life expectancy (Y)**: This is a **numerical** variable and shows the life expectancy of each country in age
2. **Country(X)**: This is a **categorical** variable and shows the countries name
3. **Year(Y): This** is a **numerical** variable and shows the year
4. **Status(X)**: This is a **categorical** variable and shows if a country is developing or developed
5. **Adult Mortality(X)**: This is a **numerical** variable and shows the Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
6. **Infant deaths(X):** This is a **numerical** variable and shows the number of Infant Deaths per 1000 population
7. **Alcohol(X): This** is a **numerical** variable and shows the recorded per capita (15+) Alcohol consumption (in liters of pure alcohol)
8. **Percentage expenditure(X):**  This is a **numerical** variable and shows the Expenditure on health as a percentage of Gross Domestic Product per capita(%)
9. **Hepatitis B(X): This** is a **numerical** variable and shows the Hepatitis B immunization coverage among 1-year-olds (%)
10. **Measles(X): This** is a **numerical** variable and shows the Measles reported cases per 1,000 population
11. **BMI(X): This** is a **numerical** variable and shows the Average Body Mass Index of entire population
12. **Under-five deaths(X): This** is a **numerical** variable and shows the number of under-five deaths per 1,000 population
13. **Polio(X): This** is a **numerical** variable and shows the Polio (Pol3) immunization coverage among 1-year-olds (%)
14. **Total expenditure(X): This** is a **numerical** variable and shows the General government expenditure on health as a percentage of total government expenditure (%)
15. **Diphtheria(X):** This is a **numerical** variable and shows the Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
16. **HIV/AIDS(X): This** is a **numerical** variable and shows the Deaths per 1,000 live births HIV/AIDS (0-4 years)
17. **GDP(X): This** is a **numerical** variable and shows the Gross Domestic Product per capita (in USD)
18. **Population(X): This** is a **numerical** variable and shows the Population of the country
19. **Thinness 1-19 years(X):** This is a **numerical** variable and shows the Prevalence of thinness among children and adolescents for Age 10 to 19 (% )
20. **Thinness 5-9 years(X): This** is a **numerical** variable and shows the Prevalence of thinness among children for Age 5 to 9(%)
21. **Income composition of resources(X): This** is a **numerical** variable and shows the Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
22. **Schooling(X): This** is a **numerical** variable and shows the number of years of Schooling

Before I began analyzing the data, I cleaned the data by dropping missing values using Python’s function called dropna(). After executing this, I was left with 1,694 rows of data. I then renamed majority of the variables to get rid of spacing errors, which in turn make it easier to call for variables when I begin running my analysis.

To begin my analysis, I decided it was appropriate to analyze a correlation matrix which would test the strength of the relationship between an X and Y variable. In this case, I was testing the strength of the relationship between the target variable (Y) “Life\_Expectancy” and all 21 variables. **See figure 1.**

**Figure 1:**

A picture containing chart

Description automatically generated

**Table 1:**

Text

Description automatically generated with medium confidence

To summarize the correlation matrix, I created a table to easily interpret the results of the relationship. Based off the correlation table (Table 1), we know that the following variables have an extreme relationship with the target variable Life\_Expectancy:

**Positive relationship:**

1. Schooling
2. Income composition of resources
3. BMI
4. GDP
5. Percentage of expenditure

**Negative relationship:**

1. Adult mortality
2. HIV/AIDS
3. Thinness 1-19 years
4. Thinness 5-9 years
5. Under 5 deaths
6. Infant deaths

The positive relationships indicate that Life\_Expectancy is likely to increase when people in these countries have more education, more income to spend on resources, a higher BMI index, a high GDP, and spend a larger percentage of their GDP on health. In contrast, the negative relationships indicate that Life\_Expectancy is likely to decrease when there’s a high rate of HIV/AIDS, deaths in children and adults, and children under 20 are too thin and do not meet the standard BMI. **See figure 2,3, 4, and 5 for visual example.**

**Figure 2:**

Chart, scatter chart

Description automatically generated

**Figure 3**

Chart, scatter chart

Description automatically generated

**Figure 4:**

Chart, scatter chart

Description automatically generated

**Figure 5:**

Chart, scatter chart

Description automatically generated

In the next section of my analysis, I created a histogram to evaluate the count life expectancies. I found that this variable is a bit left skewed which indicates that the average life expectancy of 69.3 will fall below the median and mode. In other words, I can infer that estimations of a long life exectany suggest that most people live to old age, but there are very few who survive past the average life. This brings us to our next variable: Status. Those who live in developed countries are more likely to meet or exceed the average life expectancy while those in developing countries are less likely to.

**Figure 6:**

Chart, histogram

Description automatically generated

In the final part of my analysis, I’m analyzing the descriptive statistics of different variables to understand the following:

1. Which countries meet the average life expectancy, and which do not?
2. Which country has the lowest life expectancy?
3. Which country has the highest infant deaths?
4. Which country has the highest HIV/AIDS rate as well as Measles and Hepatitis B?
5. Which status country (developed or developing) have the highest average life expectancy?

Based on my findings, countries like France, Italy, Portugal, Germany, and Belgium have the highest life expectancy of 89.0 years. I can infer their life expectancy is high most likely due to the fact that there’s universal health insurance in many of these countries. As a result, people are more likely to address and/or mitigate medical issues and have a higher chance of avoiding death. In contrast, countries like Malawi, Zimbabwe, Lesotho, and Sierra Leone have a life expectancy of less than 45 years and that’s probably a result of the impoverished living conditions and not having access to medical care. Unsurprisingly, countries considered developed under the “Status” variable have the highest average life expectancy of 78.69 years while developing countries have an average life expectancy of just 67.68 which still does not meet the world average life expectancy of at least 69.3.

Next, I wanted to evaluate countries based on their infant deaths and to my surprise, India is first on the list with 1,600 infant deaths. Next on that list is Nigeria then China, Pakistan, and Bangladesh. The top 80% of this list consists mostly of Asian countries. This could be due to the lack of basic medical care, infections, or even premature deaths and delivery complications.

Lastly, I analyzed each country and their rate of HIV/AIDS per 1,000 live births, Hepatitis B, and Measles. I was mostly focused on the HIV/AIDS section only because I know this kills more people in underdeveloped countries at an alarming rate—far more than Measles and Hepatitis B. Countries with the highest levels of HIV/AIDS include Swaziland, Zimbabwe, Botswana, Lesotho, and South Africa. In contrast, countries with the lowest count of HIV/AIDS include Afghanistan, Montenegro, Mongolia, Mexico, and Malta.

The dataset I am using is from [Kaggle](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)