Data Cleaning Report: World Life Expectancy Dataset

Objective: The goal of this report was to clean and preprocess the World Life Expectancy dataset by addressing three key data quality issues: handling duplicates, populating missing values in the "Status" column, and correcting missing life expectancy data. These actions ensure the dataset is accurate, consistent, and ready for further analysis.

1. Duplicate Records Removal:

Duplicates in the dataset were identified by checking the combination of the Country and Year columns. The following steps were taken:

- **Step 1:** A query was run to find duplicate records by grouping the data on the Country and Year columns and counting occurrences of each unique combination. Three countries were identified as having duplicates: Ireland, Senegal, and Zimbabwe.
- **Step 2:** To remove duplicates, a subquery was used to identify the Row_ID of the duplicate rows. These rows were then deleted, ensuring the dataset only retained unique records for each Country and Year.

2. Handling Missing Values in the 'Status' Column:

The Status column, which categorizes countries as either "Developing" or "Developed", had missing values that needed to be populated.

- **Step 1:** A query identified all rows where the Status column was empty.
- **Step 2:** It was found that countries marked as "Developing" already had their Status populated. A query was written to populate missing Status values with "Developing" for countries already categorized as developing.
- **Step 3:** Similarly, for rows where the Status was still missing, the Status was updated to "Developed" where applicable.

3. Handling Missing Life Expectancy Data:

There were missing values in the Life expectancy column, which needed to be populated. To address this:

- **Step 1:** A query was run to identify rows where Life expectancy was blank.
- **Step 2:** For each missing Life expectancy value, the average of the previous and next year's values was calculated to estimate the missing value for that year. This method was chosen because life expectancy typically shows gradual changes over time.

Conclusion:

After the above steps, the World Life Expectancy dataset was cleaned of duplicates, missing status values were populated, and the missing life expectancy data was accurately filled by averaging the surrounding years. The cleaned dataset is now ready for further analysis or reporting.

World Life Expectancy and Economic Correlations: An Exploratory Data Analysis

Executive Summary

This report presents an exploratory data analysis (EDA) on global life expectancy trends, utilizing the dataset world_life_expectancy. The analysis investigates several factors impacting life expectancy, including economic indicators such as GDP, BMI, and adult mortality, over a 15-year period. The findings reveal significant insights into the correlations between GDP, BMI, and life expectancy, and provide a deeper understanding of the disparities in health and economic conditions across countries.

Methodology

The analysis includes a variety of SQL queries and statistical methods to explore trends and correlations in life expectancy. Key methods used are:

- **Aggregation Functions:** To calculate minimum, maximum, and average life expectancy by country, year, and economic factors.
- **Filtering:** To exclude data points with zero values to ensure meaningful analysis.
- **Case Statements:** For categorizing GDP values to compare life expectancy between countries with high and low GDP.

- **Rolling Total:** To analyze trends in adult mortality and its relationship with life expectancy over time.

Key Findings

1. Life Expectancy Trends by Country

- **Min/Max Life Expectancy:** The analysis identified countries with extreme life expectancy values, excluding zeros. Notable trends show a significant increase in life expectancy in countries like Zimbabwe, Haiti, and Eritrea.
- **Life Expectancy Over Time:** On a global scale, life expectancy has steadily increased from 66.75 years in 2007 to 71.62 years in 2022.

2. Correlation Between GDP and Life Expectancy

- A positive correlation was found between GDP and life expectancy: higher GDP generally correlates with higher life expectancy. Countries with a higher GDP (greater than $1500) showed an average life expectancy of 74.20 years, while those with lower GDPs had an average of 64.69 years.

3. GDP Categorization and Life Expectancy

- By categorizing countries based on GDP thresholds, we observed that 1,326 countries had a GDP higher than $1500, with a corresponding average life expectancy of 74.20 years. Conversely, 1,612 countries with a GDP lower than $1500 had a life expectancy of 64.69 years.
- The life expectancy gap between countries with high and low GDP underlines the significant role that economic development plays in health outcomes.

4. Developed vs. Developing Countries

- **Life Expectancy by Country Status:** Developing countries had an average life expectancy of 66.8 years, whereas developed countries had an average of 79.2 years. This significant gap highlights the disparities between the two groups in terms of healthcare and economic conditions.

5. Life Expectancy and BMI

- A positive correlation was found between BMI and life expectancy. Countries with higher average BMI tended to have higher life expectancy, suggesting a possible link between nutritional and health factors.

6. Adult Mortality and Life Expectancy

- **Rolling Total of Adult Mortality:** A rolling total of adult mortality was used to assess the cumulative impact of mortality rates over time for countries like the United States. The analysis revealed significant fluctuations in mortality that are tied to life expectancy trends.

Conclusion

The analysis underscores the influence of economic factors such as GDP and BMI on life expectancy. It confirms that countries with higher GDPs tend to have better healthcare outcomes, as evidenced by longer life expectancies. Conversely, countries with lower GDPs face challenges in improving life expectancy, likely due to factors such as limited healthcare resources and poor living conditions.

This report also highlights the importance of adult mortality rates and nutrition (as represented by BMI) in shaping health outcomes across countries. The disparities between developed and developing countries emphasize the need for continued focus on global health initiatives, particularly in lower-income regions.