World Life Expectancy Project

Introduction

Life expectancy is one of the most critical indicators of a country's overall health, economic stability, and quality of life. This project analyzes global life expectancy trends by examining key factors such as GDP, development status, BMI, and adult mortality rates. Through data cleaning and exploratory data analysis (EDA), we ensure that our dataset is accurate and free of inconsistencies, allowing for meaningful insights into how different countries compare in terms of longevity.

By leveraging structured queries, we explore correlations between economic prosperity and life expectancy, identify trends in health and mortality rates, and uncover disparities between developed and developing nations. Our findings shed light on how strong healthcare systems, economic investment, and public health policies contribute to extended life spans, offering valuable insights for future research and policy-making efforts.

This project serves as a comprehensive examination of the global factors influencing life expectancy and highlights the importance of economic and healthcare development in improving quality of life worldwide.

Data Cleaning Process

Our dataset consists of a CSV file containing 2,941 rows of real-world data. The first steps in our workflow involve uploading the data, performing data cleaning, and conducting exploratory data analysis.
The dataset includes some missing values in the status and life expectancy columns.

Approach to Cleaning Duplicates

To verify the existence of duplicates, we focus on the Country and Year columns. Since each row should represent a unique combination of these attributes, we concatenate these columns to create a new identifier. By counting occurrences of the concatenated values, we can identify duplicates efficiently.

```
SELECT Country, Year, CONCAT(Country, Year) AS Unique_Identifier, COUNT(CONCAT(Country, Year)) A
FROM world_life_expectancy
GROUP BY Country, Year, CONCAT(Country, Year)
HAVING COUNT(CONCAT(Country, Year)) > 1;
```

This query reveals three duplicate entries in the dataset.

Removing Duplicates

To remove duplicates, we need to identify their Row_ID values (since every row has a unique identifier). By using PARTITION BY and ROW_NUMBER, we isolate duplicate entries while retaining only one instance per unique combination of Country and Year.

```
SELECT *
FROM (
    SELECT Row_ID, CONCAT(Country, Year) AS Unique_Identifier,
           ROW_NUMBER() OVER (PARTITION BY CONCAT(Country, Year) ORDER BY CONCAT(Country, Year)
    FROM world_life_expectancy
) AS Row_Table
WHERE Row_Num > 1;
```

This query helps us pinpoint duplicates for removal, ensuring the dataset remains clean and consistent for further analysis.

Removing Duplicate Entries

After identifying three duplicate rows in the dataset, the next step involves removing them. We achieve this by referencing their unique Row_ID values within a nested query. The following SQL statement

ensures these duplicates are deleted:

```
DELETE FROM world_life_expectancy
WHERE Row_ID IN (
    SELECT Row_ID
    FROM (
        SELECT Row_ID, CONCAT(Country, Year) AS Unique_Identifier,
               ROW_NUMBER() OVER (PARTITION BY CONCAT(Country, Year) ORDER BY CONCAT(Country, Ye
        FROM world_life_expectancy
    ) AS Row_Table
    WHERE Row_Num > 1
);
```

This query removes all duplicate rows from the dataset, leaving only unique Country and Year combinations.

Addressing Missing Data in the Status Column

Upon analyzing the Status column, we observe that some entries are blank or null. To address this, we take the following approach:

Identifying Blank Status Entries

We begin by checking the rows with missing values and examining the distinct existing statuses.

```
SELECT *
FROM world_life_expectancy
WHERE Status = '';

SELECT DISTINCT(Status)
FROM world_life_expectancy
WHERE Status <> '';
```

The results reveal two valid statuses: "Developing" and "Developed."

Filling Missing Statuses

To populate blank Status fields, we assign "Developing" or "Developed" based on other entries for the same country. Here's the logic:

• If a country has rows marked as "Developing," any blank statuses for that country should also be filled with "Developing."

• Similarly, blank statuses are filled with "Developed" if the country has entries marked as "Developed."

Initial attempt using UPDATE with subqueries:

```
UPDATE world_life_expectancy
SET Status = 'Developing'
WHERE Country IN (
    SELECT DISTINCT(Country)
    FROM world_life_expectancy
    WHERE Status = 'Developing'
);
```

However, this approach doesn't work as expected. To resolve the issue, we use a self-join in the UPDATE statement:

```
UPDATE world_life_expectancy t1
JOIN world_life_expectancy t2
    ON t1.Country = t2.Country
SET t1.Status = 'Developing'
WHERE t1.Status = ''
  AND t2.Status = 'Developing';
```

The same logic applies for filling "Developed" statuses:

```
UPDATE world_life_expectancy t1
JOIN world_life_expectancy t2
    ON t1.Country = t2.Country
SET t1.Status = 'Developed'
WHERE t1.Status = ''
  AND t2.Status = 'Developed';
```

With these updates, all blank statuses are now populated appropriately.

Addressing Missing Data in the Life Expectancy Column

Our next step involves checking the Life expectancy column for missing values. Upon analysis, we discovered two null entries in this column. To handle these blanks, we decided to populate them with the average of the life expectancy values from the year immediately before and the year immediately after for the same country. This approach is deemed accurate due to the limited number of null entries and the consistency of life expectancy trends.

Identify Missing Values

We begin by identifying the rows where the Life expectancy column is null. This allows us to locate the specific countries and years affected.

```
SELECT *
FROM world_life_expectancy
WHERE `Life expectancy` = '';
SELECT Country, Year, `Life expectancy`
FROM world_life_expectancy
WHERE `Life expectancy` = '';
```

Calculate Averages for Missing Values

Next, we calculate the average life expectancy using the values from one year prior and one year after the missing entry for the same country. To achieve this, we perform a join operation that links the data for these consecutive years.

```
SELECT t1.Country, t1.Year, t1.`Life expectancy`,
       t2.Country, t2.Year, t2.`Life expectancy`,
       t3.Country, t3.Year, t3.`Life expectancy`,
       ROUND((t1.`Life expectancy` + t2.`Life expectancy`) / 2, 1)
FROM world_life_expectancy t1
JOIN world_life_expectancy t2
    ON t1.Country = t2.Country
    AND t1.Year = t2.Year - 1
JOIN world_life_expectancy t3
    ON t1.Country = t3.Country
    AND t1.Year = t3.Year + 1
WHERE t1.`Life expectancy` = '';
```

Update Missing Values

Using the calculated averages, we update the null entries in the Life expectancy column. The following query ensures that the missing values are replaced with the computed averages.

```
UPDATE world_life_expectancy t1
JOIN world_life_expectancy t2
    ON t1.Country = t2.Country
    AND t1.Year = t2.Year - 1
JOIN world_life_expectancy t3
    ON t1.Country = t3.Country
    AND t1.Year = t3.Year + 1
SET t1.`Life expectancy` = ROUND((t2.`Life expectancy` + t3.`Life expectancy`) / 2, 1)
WHERE t1.`Life expectancy` = '';
```

By implementing this process, the two null values in the Life expectancy column have been replaced with accurate averages derived from neighboring data points. This ensures a complete and consistent dataset for further analysis.

Data Analysis

Life Expectancy Trends Over 15 Years

In this section, we analyze the Life expectancy column to evaluate how each country has progressed over the last 15 years. Our focus is on identifying both the lowest and highest life expectancy values for each country during this period.
Some entries contain zero values, which could distort results, so we use HAVING to filter them out.
The goal is to determine which countries have made the most significant progress by calculating the difference between maximum and minimum life expectancy values.
Countries such as Haiti, Zimbabwe, and Eritrea have demonstrated remarkable improvements, increasing their life expectancy by 28 years in just 15 years.

```
SELECT Country,
       MIN(`Life expectancy`) AS Min_Life_Exp,
       MAX(`Life expectancy`) AS Max_Life_Exp,
       ROUND(MAX(`Life expectancy`) - MIN(`Life expectancy`), 1) AS Life_Increase_15_Years
FROM world_life_expectancy
GROUP BY Country
HAVING MIN(`Life expectancy`) <> 0
   AND MAX(`Life expectancy`) <> 0
ORDER BY Life_Increase_15_Years DESC;
```

This query extracts the minimum and maximum life expectancy for each country and calculates the improvement over the analyzed period. Sorting by Life_Increase_15_Years allows us to identify countries with the highest gains in life expectancy.

GDP vs. Life Expectancy Analysis

We analyze the relationship between GDP and Life expectancy to determine how economic status influences health outcomes. To ensure accuracy, we filter out rows where either GDP or Life expectancy equals zero.
Observations reveal that lower GDP nations tend to have lower life expectancy rates due to limited healthcare infrastructure and resources. In contrast, high GDP countries, such as Switzerland, Luxembourg, and Qatar, exhibit life expectancy levels 20 to 30 years higher than nations with lower GDPs—typically around 80 years.
This indicates a positive correlation between GDP and life expectancy.

```
SELECT Country, ROUND(AVG(`Life expectancy`),1) AS Life_Exp, ROUND(AVG(GDP),1) AS GDP
FROM world_life_expectancy
GROUP BY Country
HAVING Life_Exp > 0 AND GDP > 0
ORDER BY GDP DESC;
```

Categorizing Countries by GDP

To better analyze trends, we categorize countries based on GDP values using a CASE statement. We define 1,500 GDP as the midpoint, based on our dataset:
• High GDP Countries: GDP ≥ 1,500
• Low GDP Countries: GDP < 1,500

```
SELECT
SUM(CASE WHEN GDP >= 1500 THEN 1 ELSE 0 END) AS High_GDP_Count
FROM world_life_expectancy;
```

From this, we find 1,326 rows categorized as high GDP nations.
Comparing Life Expectancy in High vs. Low GDP Countries
To refine the analysis, we calculate the average life expectancy for both groups.

```
SELECT
SUM(CASE WHEN GDP >= 1500 THEN 1 ELSE 0 END) AS High_GDP_Count,
AVG(CASE WHEN GDP >= 1500 THEN `Life expectancy` ELSE NULL END) AS High_GDP_Life_expectancy,
SUM(CASE WHEN GDP <= 1500 THEN 1 ELSE 0 END) AS Low_GDP_Count,
AVG(CASE WHEN GDP <= 1500 THEN `Life expectancy` ELSE NULL END) AS Low_GDP_Life_expectancy
FROM world_life_expectancy;
```

We can see that countries with higher GDP levels tend to have longer life expectancy, with an average of 75 years in high GDP nations compared to 65 years in low GDP nations. This 10-year difference highlights the strong correlation between economic prosperity and public health, as wealthier countries generally have better healthcare infrastructure, medical access, and living conditions, contributing to increased longevity.

Life Expectancy by Development Status

Another important factor to analyze is status, which categorizes countries as either Developing or Developed. To compare life expectancy between these groups, we calculate the average life expectancy for each status:

```
SELECT Status, ROUND(AVG(`Life expectancy`),1) AS Avg_Life_Expectancy
FROM world_life_expectancy
GROUP BY Status;
```

The results show that:
• Developing countries have an average life expectancy of 67 years.
• Developed countries have an average life expectancy of 80 years.
However, this comparison does not give us a complete picture, as the distribution of countries across statuses can significantly impact the average values. If there are many developing countries and only a few developed ones, the overall average might be skewed downward.
To understand the distribution of countries within each category, we run the following query:

```
SELECT Status, COUNT(DISTINCT Country) AS Country_Count, ROUND(AVG(`Life expectancy`),1) AS Avg_
FROM world_life_expectancy
GROUP BY Status;
```

The results show that:
• 32 developed countries are included in the dataset.
• 161 developing countries are represented.
This distribution confirms that the higher number of developing countries lowers the overall average, potentially skewing the comparison between the two groups.

BMI vs. Life Expectancy

Another important factor to examine is BMI (Body Mass Index), which can have a direct impact on life expectancy. To analyze this relationship, we calculate the average BMI and life expectancy for each country, filtering out rows where either value is zero to ensure accuracy.
The query below shows the average Life expectancy by BMI.

```sql
SELECT Country, ROUND(AVG(`Life expectancy`),1) AS Life_Exp, ROUND(AVG(BMI),1) AS BMI
FROM world_life_expectancy
GROUP BY Country
HAVING Life_Exp > 0 AND BMI > 0
ORDER BY BMI DESC;
```

Research indicates that higher BMI is associated with lower life expectancy, as increased BMI can lead to higher risks of heart disease, diabetes, and premature death. Countries with higher average BMI levels tend to have shorter lifespans, reinforcing the negative health impact of obesity.

Tracking Adult Mortality Trends

The Adult Mortality column provides insight into how many individuals die each year in each country. To observe mortality trends over time, we compute a rolling sum of adult mortality for each country using the PARTITION BY function. The query below shows rolling adult mortality over Time.

```sql
SELECT Country, Year, `Life expectancy`, `Adult Mortality`,
       SUM(`Adult Mortality`) OVER (PARTITION BY Country ORDER BY Year) AS Rolling_Total
FROM world_life_expectancy
WHERE Country LIKE '%United%';
```

For example, in the United States, the rolling total starts at 114 and increases to 931 over time, indicating a rise in adult mortality rates. Tracking this metric helps identify patterns in public health crises, disease outbreaks, and shifts in healthcare effectiveness.

Conclusion

Through data cleaning and exploratory analysis, this project has provided valuable insights into global life expectancy trends. By removing duplicates and addressing missing values, we ensured a clean dataset for accurate analysis. Our findings reveal significant disparities in life expectancy across countries, largely influenced by factors such as economic status (GDP), development classification, BMI, and adult mortality rates.

One of the most striking observations is the strong correlation between GDP and life expectancy—countries with higher GDP levels tend to have longer life expectancies, with a difference of up to 10 years between high and low GDP nations. Similarly, developed countries have an average life expectancy of 80 years, compared to 67 years in developing nations, though the distribution of countries in each group skews the overall averages.

Additional health-related factors such as BMI and adult mortality trends further demonstrate the impact of health infrastructure, economic investment, and disease prevention on longevity. Countries with higher BMI averages show lower life expectancy due to higher risks of heart disease and other health complications, reinforcing the importance of public health measures.

Overall, this project highlights how socioeconomic and health factors shape life expectancy trends worldwide. These insights can help guide policies aimed at improving healthcare access, economic stability, and public health initiatives, ultimately contributing to longer, healthier lives across nations.