ILS Z511 DATABASE DESIGN

**FINAL PROJECT REPORT**

**Global Economic Development Analysis**

**Nikhil Sai Jaddu**

**Anusha Chebolu**

**Shalini Cherukuri**

**Kundhana Harshitha Paruchuru**

# INTRODUCTION

In an increasingly interconnected global economy, comprehensive data is essential for policymakers to make informed decisions that promote sustainable economic growth and social welfare. This project provides **economists and government policymakers** with a detailed dataset that includes a variety of key indicators. Sourced from authoritative institutions including The World Bank, World Happiness Report and Transparency International, the dataset offers crucial insights into the economic performance and well- being of countries worldwide.

# COMMUNITY OR POPULATION

The target community for this project includes **economists and policymakers within government agencies**. These professionals utilize data to assess economic performance, labor force dynamics, and infrastructure access, enabling evidence-based policy decisions that promote sustainable development. This dataset will support their efforts in identifying development challenges and crafting effective interventions to enhance national economic stability and social well-being.

## Why we chose the community/population

We have identified **economists and policy makers** as the target community for this project for several key reasons:

1. **Macro-Level Influence**: Economists play a pivotal role in advising governments and organizations on the efficient allocation of resources, directly impacting economic outcomes and societal well-being.
2. **Global Relevance**: Economic indicators such as labor market trends, GDP growth, and corruption perception have both national and international significance. Economists are instrumental in shaping policies that address these issues at a global scale, ensuring the alignment of national strategies with international economic standards.
3. **Comprehensive Impact**: Economists inform critical decisions related to national employment strategies, economic growth, and anti-corruption efforts. By providing accurate and comprehensive data, this project supports economists in formulating policies aimed at reducing poverty, creating sustainable employment opportunities, and promoting long-term economic development.

This focus on economists ensures that the data provided will have a wide-reaching influence, driving informed decision-making across both national and global platforms.

# DATASET DESCRIPTION

The dataset is a compilation of World Development Indicators from 2015 to 2018. The dataset has been sourced from three primary, reputable organizations. The World Bank provides data on essential economic and social variables, including GDP, GDP per capita, population, pollution levels, life expectancy, access to electricity, and labor force participation rates. The world happiness Report contains social variables such as social support, generosity, and the freedom to make life choices. Transparency International offers data regarding corruption perception, measured by the Corruption Perceptions Index (CPI).

## Variables:

The dataset contains 186 rows (representing 186 countries), and 20 columns sourced from 2015-18.

|  |  |
| --- | --- |
| country | Name of the country |
| electricity\_access | Access to electricity (% of population) |
| gdp | GDP (current US$) |
| gdp\_capita | GDP per capita (current US$) |
| labor\_rate | Labor force participation rate, total (% of total population ages 15+) |
| labor\_force | Labor force, total |
| land\_area | Land area (sq. km) |
| life\_expectancy | Life expectancy at birth, total (years) |
| adult\_literacy | Literacy rate, adult total (% of people ages 15 and above) |

|  |  |
| --- | --- |
| water\_access | People using at least basic drinking water services (% of population) |
| air\_pollution | PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total) |
| population\_density | Population density (people per sq. km of land area) |
| population | Population, total |
| alcohol\_consumption | Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age) |
| unemployment\_rate | Unemployment, total (% of total labor force) (modeled ILO estimate) |
| social\_support | Social Support rating based on World Happiness Report |
| freedom | Freedom to make life choices rating based on World Happiness Report |
| generosity | Generosity rating based on World Happiness Report |
| income\_class | Income Classification |
| cpi | Corruption Perceptions Index |

# How the dataset(s) you have chosen are related to the community/population you have chosen.

The dataset World Development Data is directly relevant to the community of **economists** because it includes critical indicators that economists use to evaluate the health and progress of national economies. For instance, the dataset provides metrics like GDP, GDP per capita, labor force participation rates, unemployment rates, and income classifications, all of which are fundamental to understanding a country's economic performance. Economists rely on this data to assess economic growth, productivity, and income distribution, enabling them to forecast economic trends and provide informed recommendations to policymakers.

In particular, the dataset's focus on both **economic** and **social factors** (such as labor, social support, and public health data) allows economists to study the broader societal impacts of economic policies. For example, analyzing the correlation between GDP per capita and indicators like life expectancy, unemployment rate or social support can help economists understand the extent to which economic growth translates into improved living standards. Additionally, metrics like the Corruption Perceptions Index enable economists to explore how governance and transparency affect economic stability and development. This comprehensive dataset aligns with economists' need for multidimensional analysis of economic systems and their socio-political contexts, making it an invaluable resource for shaping economic policies that promote sustainable development.

# PROBLEMS TO BE SOLVED

### How can labor force indicators be used to identify labor market health?

The project aims to address labor market challenges by analyzing key indicators such as labor force participation rate and labor force size. These indicators provide critical insights into global labor dynamics, enabling policymakers to identify areas for improvement and design targeted interventions to foster economic resilience.

### Labor Force Participation Rate:

This refers to the percentage of the working-age population (15-64) employed or actively seeking employment. A low participation rate indicates underutilized potential. Conversely, a high participation rate reflects an engaged workforce but may also signal insufficient job creation if accompanied by high unemployment.

**Implications**: Countries with low participation may need to improve access to education, skills training, and gender inclusion. High-participation nations must focus on creating jobs to reduce unemployment and getting global companies to invest and uplift the job market.

### Labor Force Size:

This refers to the total number of people employed or seeking employment. A growing labor force can drive economic growth, but without adequate job opportunities, it risks leading to unemployment and social instability. A shrinking labor force may result in labor shortages and slower economic growth.

**Implications**: Growing economies must prioritize job creation, while countries facing labor shortages should consider immigration policies, workforce planning, and automation or productivity improvements.

**REASON FOR USAGE:** This project focuses on analyzing several indicators that affect a country’s economy, some of the significant factors are labor indicators was chosen to explore relationship between economy and labor market. This question aligns directly with the role of economists in identifying weaknesses in labor markets and providing solutions for sustainable economic growth.

### Does a higher GDP per capita lead to increased social support and freedom to make life choices?

GDP per capita is often used as a measure of a country’s economic prosperity, but economists are also interested in how this wealth translates into improved social conditions. By analyzing the relationship between GDP per capita and social indicators like social support and freedom to make life choices, economists can assess the broader impact of economic growth on societal well-being.

**Implications**: A country experiencing GDP growth but facing low literacy rates may prioritize educational initiatives to develop a more skilled workforce. Furthermore, implementation of social support systems, including mental health services and community engagement programs ensures that economic growth translates into meaningful benefits for its citizens.

**REASON FOR USAGE:** The question was chosen to explore the potential link between economic prosperity (GDP per capita) and societal well-being (social support, freedom). This question allows economists to provide evidence on whether economic prosperity alone is sufficient for societal well-being or if additional social policies are required.

### What is the correlation between corruption perception and income class across countries?

Corruption Perception is typically assessed using the Corruption Perception Index (CPI), which assigns values ranging from 0 to 100, where higher scores indicate lower perceived corruption, and lower scores reflect higher levels of corruption. Countries are commonly classified into distinct income categories (low, lower-middle, upper-middle, high income) based on their GDP per capita. This project will examine the relationship between income categories and corruption perception by exploring how different countries, across these income classifications, exhibit varying levels of perceived corruption. Countries with higher levels of corruption (indicated by lower CPI scores) may experience slower economic growth, driven by factors such as inefficiencies in governance, misallocation of resources, and diminished investor confidence. In contrast, countries with lower levels of corruption (higher CPI scores) are often characterized by stronger economic performance, as reduced corruption fosters a more favorable environment for investment, resource allocation, and efficient governance.

**Implications:** Policy makers can reduce corruption and enhance economic outcomes by promoting transparency, protecting whistleblowers, and conducting public awareness campaigns. Additionally, judicial reforms and regular performance evaluations can further improve accountability and efficiency, fostering a stable environment for economic growth.

**REASON FOR USAGE**: This question was chosen to examine the relationship between income categories and corruption perception. This question is critical for economists because it addresses the broader issue of governance and its impact on economic performance, which is essential for formulating policies aimed at reducing corruption and fostering inclusive growth.

### How can health-related indicators be used to assess public health and environmental sustainability?

Health-related indicators, such as life expectancy, alcohol consumption and air pollution, provide critical insights into the overall health of populations. These metrics, when combined, enable an understanding of the interconnectedness between public health and environmental conditions. This project will examine how air pollution levels influence life expectancy and how alcohol consumption interacts with these dynamics. High air pollution levels are often correlated with reduced life expectancy due to their association with chronic health conditions such as asthma, cancer, and cardiovascular diseases. On the other hand, alcohol consumption may serve as an indicator of lifestyle and socioeconomic factors, which could further amplify or mitigate health outcomes. By comparing these indicators across countries with varying environmental policies and health systems, the analysis can uncover patterns that link environmental degradation, public health, and social behaviors.

**Implications:** Policymakers can leverage these insights to design targeted interventions. For example, reducing air pollution through stricter environmental regulations can directly enhance life expectancy. Simultaneously, public health campaigns aimed at reducing excessive alcohol consumption can mitigate its impact on overall health outcomes. Integrated policies that address both environmental and lifestyle factors can significantly improve public health and sustainability outcomes.

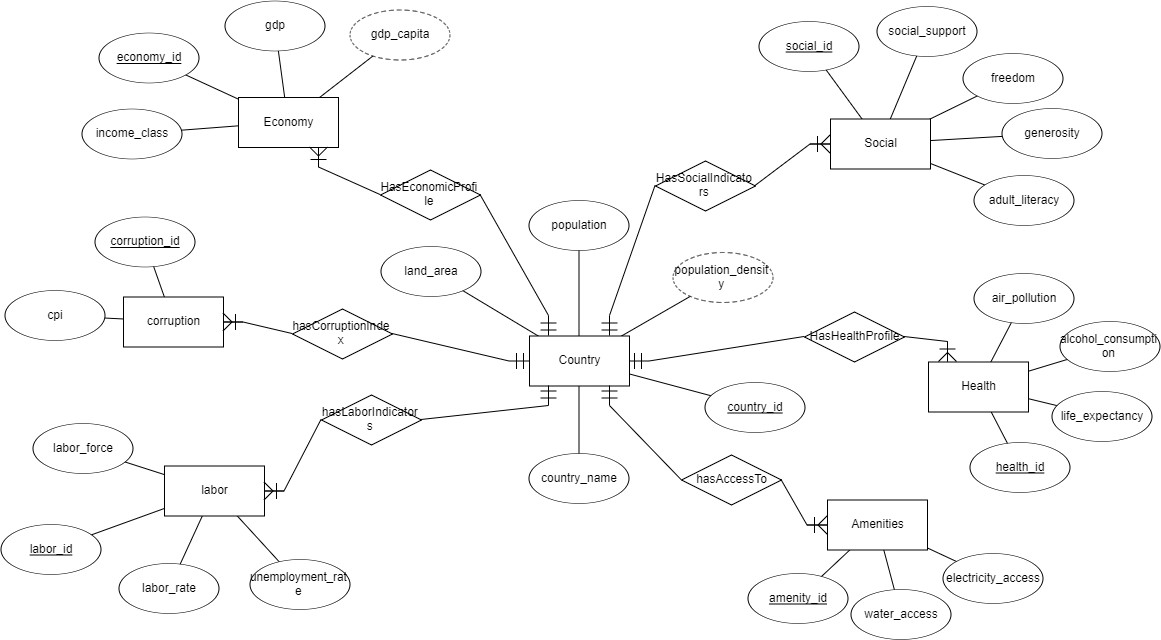
**REASON FOR USAGE**: This question was chosen to investigate the interplay between environmental quality (air pollution), public health outcomes (life expectancy), and lifestyle factors (alcohol consumption). Understanding these connections is essential for crafting policies that simultaneously address environmental and public health challenges, contributing to sustainable development and improved quality of life.

# CHEN ER DIAGRAM

The dataset comprises 20 columns, categorized into 7 entities that reflect the geographic, social, economic, health, labor, and infrastructure aspects of a nation. Below is a sample row from the dataset. The relationships between the entities are shown in the ER diagram below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| countr y | electricity  \_access | gd p | gdp\_ capit a | labor  \_rate | labor\_forc e | land\_are a | life\_expectanc y | adult\_literac y | water\_acce ss |
| Afghan | 90.08 | 1.8 | 536.2 | 47.28 | 9103246 | 652230 | 62.97 |  | 66.61 |
| istan |  | 8E | 3 |  |  |  |  |  |
|  |  | +1 |  |  |  |  |  |  |
|  |  | 0 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| air\_po llution | populatio n\_density | population | alcohol  \_consu mption | unemployme nt\_rate | social\_ suppor t | freedom | generosity | income\_ class | cpi |
| 100 | 53.94 | 35179977 | 0.0087 | 10.341 | 0.522 | 0.428 | -0.014 | Low income | 14.25 |



## NORMALISATION

## Normalization in this project was achieved by designing a database schema that adheres to higher normal forms (1NF, 2NF, and 3NF) to reduce redundancy and dependency issues.

## Breaking Down Data into Logical Tables:

## The data was divided into multiple related tables such as:

## Country: Stores country-specific static information (e.g., name, population, land area).

## Economy: Stores economic indicators (e.g., GDP, income class).

## Health: Contains health-related data (e.g., life expectancy, air pollution, alcohol consumption).

## Social: Includes social indicators (e.g., freedom, social support).

## Corruption: Tracks corruption perception index (CPI).

## Amenities: Tracks water and electricity access.

## Labor: Labor market indicators (e.g., labor rate and unemployment).

## Avoiding Redundancy:

## Each attribute is stored in only one table. For example, the country name is stored in the Country table, avoiding repetition in other tables. Instead, foreign keys linked other tables to the Country table using a unique identifier (country\_id).

## Functional Dependency:

## Attributes in each table depend only on the primary key. For example, In the Economy table, gdp and income\_class depend solely on country\_id. In the Health table, indicators like life\_expectancy and air\_pollution depend on country\_id.

## Minimizing Update and Deletion Anomalies:

## With the separation of data into specific tables, updates or deletions in one aspect of the data (e.g., updating population in Country) did not impact unrelated data (e.g., health or economic indicators).

## Using Foreign Keys for Relationships:

## Relationships between tables were established using foreign keys, such as the country\_id linking Country with Economy, Health, and others. This ensured referential integrity and seamless data management

## KEY RESEARCH QUESTIONS AND FINDINGS

### Labor Market Health

How Can Labor Force Indicators Be Used to Identify Labor Market Health?

**Objective:**

This analysis seeks to evaluate labor market health by identifying countries with labor force participation rates below specific thresholds based on the global average and standard deviation. The goal is to understand patterns of workforce engagement and identify areas requiring intervention.

**Query Design**

The SQL query identifies countries with labor force participation rates below average by applying a statistical threshold. The steps are:

1. Join Operation: The Labor table is joined with the Country table to associate labor participation data with respective country names.
2. Threshold Calculation: The global average labor force participation rate (AVG) , unemployment rate (AVG) and standard deviations (Stddev) are calculated from the Labor table.
3. Filtering: Countries with labor participation rates below and unemployment above these thresholds are identified and ordered in descending order of their labor rates for analysis.

**Key Query 1**

SELECT c.name, l.labor\_rate  
FROM Labor l  
JOIN Country c ON l.country\_id = c.country\_id  
WHERE l.labor\_rate < (SELECT AVG(labor\_rate) - n\*STDDEV(labor\_rate) FROM Labor)  
ORDER BY l.labor\_rate DESC;

* Inputs:
  + n = 1: Identifies countries below 1 standard deviation.
  + n = 2: Identifies countries below 2 standard deviations.
* Statistical Results:
  + Average labor rate (Avg): 60.7%
  + Standard deviation (StdDev): 10.11%

**Key Query 2**

*SELECT c.name, l.unemployment\_rate*

*FROM Labor l*

*JOIN Country c ON l.country\_id = c.country\_id*

*WHERE l.unemployment\_rate > (SELECT AVG(unemployment\_rate) + 1.5\*STDDEV(unemployment\_rate) FROM Labor)*

*ORDER BY l.unemployment\_rate DESC;*

* Inputs:
  + n = 1: Identifies countries above 1.5 standard deviation.
* Statistical Results:
  + Average Unemployment rate (Avg): 7.614%
  + Standard deviation (StdDev): 5.564%

**Key Query 3**

*SELECT c.name, l.unemployment\_rate , s.adult\_literacy*

*FROM Labor l join Country c join Social s ON*

*c.country\_id=l.country\_id and l.country\_id=s.country\_id*

*AND l.unemployment\_rate IS NOT NULL*

*ORDER by unemployment\_rate;*

**Findings**

1. Query 1 (n=1, n=2):
   1. Countries with labor rates below 49.55%.
   2. Result: 28 countries fall below this threshold.
   3. Countries with labor rates below 41%.
   4. Result: 6 countries fall below this stricter threshold.
2. Query 2:
   1. Countries with unemployment rates above 16%.
   2. Result: 19 countries fall above this threshold.
3. Query 3:
   1. Result: 182 countries are returned with very few nulls in literacy rate.
   2. Correlation is plotted below.

A graph with green and white dots

Description automatically generated

**Implications**

* Low Participation Rates:
  + Countries with participation rates below the global average face significant challenges, such as gender disparities, skill mismatches, and inadequate access to education or training.

**Policy Recommendations**

* + Develop and implement inclusive workforce policies to encourage higher participation.
  + Promote vocational training and education, particularly for underrepresented groups.
  + Incentives for Private Sector Hiring, Support for Elderly and Disabled Workers, Microloans and Grants for small businesses and startups
  + Industry Collaboration, Skill alignment, Invest in Emerging Sectors, Attract Foreign Investment, Support Labor Mobility, Boost Small and Medium Enterprises (SMEs) by providing tax benefits, subsidies.

### GDP and Social Well-Being

Does a Higher GDP Per Capita Lead to Increased Social Support and Freedom to Make Life Choices?

**Objective**

The goal of this analysis is to explore whether an increase in GDP per capita correlates with enhanced social support and greater freedom to make life choices. Both of these indicators are critical in assessing societal well-being and understanding how economic prosperity translates into improved quality of life.

**Query Design**

The analysis utilizes a structured SQL query to classify countries into GDP categories - Low, Medium, and High, based on their GDP per capita. These categories are determined using statistical measures, specifically the mean and standard deviation (stddev), calculated from the dataset.

1. GDP Categories:
   1. Low GDP: Countries with GDP per capita below

Mean − StdDev

* 1. Medium GDP: Countries with GDP per capita between

Mean−StdDev and Mean+StdDev.

* 1. High GDP: Countries with GDP per capita above

Mean+StdDev

1. Data Aggregation:
   1. The average values of social\_support and freedom were calculated for each GDP category using a join operation between economic and social datasets. Null values were excluded to ensure data accuracy.
2. Statistical Output:
   1. The results include the mean GDP per capita and its standard deviation, along with the average social support and freedom ratings for each GDP category.

*SELECT*

*CASE*

*WHEN e.gdp\_per\_capita < (SELECT AVG(gdp\_per\_capita) - STDDEV(gdp\_per\_capita) FROM EconomyWithGDPPerCapita WHERE gdp\_per\_capita IS NOT NULL) THEN 'Low GDP'*

*WHEN e.gdp\_per\_capita BETWEEN (SELECT AVG(gdp\_per\_capita) - STDDEV(gdp\_per\_capita) FROM EconomyWithGDPPerCapita WHERE gdp\_per\_capita IS NOT NULL)*

*AND (SELECT AVG(gdp\_per\_capita) + STDDEV(gdp\_per\_capita) FROM EconomyWithGDPPerCapita WHERE gdp\_per\_capita IS NOT NULL) THEN 'Medium GDP'*

*WHEN e.gdp\_per\_capita > (SELECT AVG(gdp\_per\_capita) + STDDEV(gdp\_per\_capita) FROM EconomyWithGDPPerCapita WHERE gdp\_per\_capita IS NOT NULL) THEN 'High GDP'*

*ELSE 'Unknown' -- To handle cases where gdp\_per\_capita is null*

*END AS gdp\_category,*

*AVG(s.social\_support) AS avg\_social\_support,*

*AVG(s.freedom) AS avg\_freedom*

*FROM EconomyWithGDPPerCapita e*

*JOIN Social s ON e.country\_id = s.country\_id*

*WHERE e.gdp\_per\_capita IS NOT NULL*

*AND s.social\_support IS NOT NULL*

*AND s.freedom IS NOT NULL*

*GROUP BY gdp\_category;*

**Findings**

1. GDP Categorization:
   1. Average GDP per capita: $13,598.81 (mean) with a standard deviation of $18,598.15.
   2. Medium GDP: Countries with GDP per capita in the range

$13,598.81 ± 18,598.15

* 1. High GDP: Countries exceeding the upper threshold of this range.

1. Correlation with Social Support:
   1. Medium GDP countries: Average social support rating of 0.7761.
   2. High GDP countries: Average social support rating of 0.9212.
   3. Insight: A noticeable increase in social support is observed as GDP per capita rises.
2. Correlation with Freedom to Make Life Choices:
   1. Medium GDP countries: Average freedom rating of 0.7485.
   2. High GDP countries: Average freedom rating of 0.8895.
   3. Insight: Higher GDP categories consistently exhibit greater freedom ratings, highlighting a strong correlation.

A screenshot of a computer

Description automatically generatedA graph of a social support

Description automatically generated

A screenshot of a social media survey

Description automatically generated

**Implications**

Countries aiming to boost GDP should complement economic strategies with targeted social programs to enhance communal support systems. Medium-GDP nations can focus on governance reforms to bridge the gap in freedom-related metrics.

**Policy Recommendations**

* 1. Supporting innovation to drive growth and development,
  2. Protecting civil liberties to ensure individual rights and freedoms
  3. Promoting work-life balance to improve well-being and productivity.

### Governance and Income Class

What is the Correlation Between Corruption Perception and Income Class Across Countries?

**Objective**

This analysis examines the relationship between corruption perception, as measured by the Corruption Perception Index (CPI), and income classes (Low, Lower Middle, Upper Middle, and High Income) across countries. The goal is to identify how income levels influence perceptions of corruption globally.

**Query Design**

The SQL query aggregates corruption perception data by income class and calculates the following statistics:

1. Average CPI (avg\_cpi): Reflects the typical perception of corruption for each income class.
2. Minimum CPI (min\_cpi): The lowest corruption perception score in each income class.
3. Maximum CPI (max\_cpi): The highest corruption perception score in each income class.
4. Number of Countries (num\_countries): Total countries within each income class.

SELECT e.income\_class,  
 AVG(c.cpi) AS avg\_cpi,  
 MIN(c.cpi) AS min\_cpi,  
 MAX(c.cpi) AS max\_cpi,  
 COUNT(c.cpi) AS num\_countries  
 FROM Economy e  
 JOIN Corruption c ON e.country\_id = c.country\_id  
 GROUP BY e.income\_class  
 ORDER BY avg\_cpi DESC;

**Findings**

1. High-Income Countries:
   1. Average CPI: 65.04
   2. Range: 34.5–89.25
   3. Number of Countries: 54
   4. Insight: High-income countries typically have lower corruption levels, as reflected in their higher CPI scores.
2. Upper Middle-Income Countries:
   1. Average CPI: 38.68
   2. Range: 16–61.25
   3. Number of Countries: 47
   4. Insight: Corruption levels increase in this group, indicating governance challenges despite economic growth.
3. Lower Middle-Income Countries:
   1. Average CPI: 33.14
   2. Range: 17.75–66.25
   3. Number of Countries: 51
   4. Insight: Widespread corruption limits development and trust in governance.
4. Low-Income Countries:
   1. Average CPI: 25.52
   2. Range: 9.25–54.75
   3. Number of Countries: 24
   4. Insight: These countries face the highest levels of perceived corruption, likely due to weak governance structures and lack of transparency.

**Implications**

1. Positive Correlation: Higher income classes are associated with higher CPI scores, indicating lower corruption perception. This trend highlights the role of economic stability in fostering stronger governance and transparency.
2. Systemic Challenges in Lower-Income Groups: Low-income and lower middle-income countries exhibit significantly lower average CPI scores, highlighting systemic corruption as a major barrier to development.

**Policy Recommendations**

1. Strengthen Governance:
   1. Invest in institutional reforms and anti-corruption frameworks, particularly in low-income and lower middle-income countries.
2. Promote Transparency:
   1. Encourage transparency in public procurement, resource allocation, and government activities to build trust.
3. Capacity Building:
   1. Develop stronger legal systems, whistleblower protections, and independent oversight mechanisms to reduce corruption.

### Public Health and Environmental Sustainability

How Can Health-Related Indicators Be Used to Assess Public Health and Environmental Sustainability?

**Objective**

This analysis examines how key health-related indicators—Life Expectancy, Air Pollution, and Alcohol Consumption—can be utilized to evaluate public health and environmental sustainability across countries. By categorizing countries based on these indicators, this analysis helps policymakers identify patterns and prioritize interventions for improving health outcomes and ensuring environmental sustainability.

**Query Design**

The analysis categorizes countries using the following indicators:

1. Life Expectancy:
   1. Categories:
      1. Low Life Expectancy: Below Mean - StdDev
      2. Medium Life Expectancy: Within Mean ± StdDev.
      3. High Life Expectancy: Above Mean + StdDev
2. Air Pollution:
   1. Categories:
      1. Green: ≤25 (Good air quality).
      2. Yellow: 26 - 60 (Moderate air pollution).
      3. Orange: 61 - 96 (Unhealthy for sensitive groups).
      4. Red: 97 - 150 (Unhealthy air quality).
      5. Purple: > 150 (Very unhealthy or hazardous).
3. Alcohol Consumption:
   1. Categories:
      1. Low Alcohol Consumption: Below Mean−StdDev
      2. Medium Alcohol Consumption: Within Mean ± StdDev .
      3. High Alcohol Consumption: Above Mean + StdDev .

*create view health\_Stata as*

*SELECT*

*c.country\_id, c.name,*

*-- Life Expectancy Categories*

*CASE*

*WHEN h.life\_expectancy < (avg\_life\_expectancy - stddev\_life\_expectancy) THEN 'Low Life Expectancy'*

*WHEN h.life\_expectancy BETWEEN (avg\_life\_expectancy - stddev\_life\_expectancy)*

*AND (avg\_life\_expectancy + stddev\_life\_expectancy) THEN 'Medium Life Expectancy'*

*WHEN h.life\_expectancy > (avg\_life\_expectancy + stddev\_life\_expectancy) THEN 'High Life Expectancy'*

*ELSE 'Unknown'*

*END AS life\_expectancy\_category,*

*-- Air Pollution Categories*

*CASE*

*WHEN air\_pollution <= 25 THEN 'Green'*

*WHEN air\_pollution BETWEEN 25 AND 60 THEN 'Yellow'*

*WHEN air\_pollution BETWEEN 60 AND 96 THEN 'Orange'*

*WHEN air\_pollution BETWEEN 98 AND 150.4 THEN 'Red'*

*ELSE 'Purple'*

*END AS air\_pollution\_category,*

*-- Alcohol Consumption Categories*

*CASE*

*WHEN h.alcohol\_consumption < (avg\_alcohol\_consumption - stddev\_alcohol\_consumption) THEN 'Low Alcohol Consumption'*

*WHEN h.alcohol\_consumption BETWEEN (avg\_alcohol\_consumption - stddev\_alcohol\_consumption)*

*AND (avg\_alcohol\_consumption + stddev\_alcohol\_consumption) THEN 'Medium Alcohol Consumption'*

*WHEN h.alcohol\_consumption > (avg\_alcohol\_consumption + stddev\_alcohol\_consumption) THEN 'High Alcohol Consumption'*

*ELSE 'Unknown'*

*END AS alcohol\_consumption\_category*

*FROM Country c JOIN*

*Health h ON c.country\_id = h.country\_id*

*-- Calculate the averages and standard deviations for health indicators*

*JOIN ( SELECT*

*AVG(h.life\_expectancy) AS avg\_life\_expectancy,*

*STDDEV(h.life\_expectancy) AS stddev\_life\_expectancy,*

*AVG(h.air\_pollution) AS avg\_air\_pollution,*

*STDDEV(h.air\_pollution) AS stddev\_air\_pollution,*

*AVG(h.alcohol\_consumption) AS avg\_alcohol\_consumption,*

*STDDEV(h.alcohol\_consumption) AS stddev\_alcohol\_consumption*

*FROM Health h WHERE*

*h.life\_expectancy IS NOT NULL*

*AND h.air\_pollution IS NOT NULL*

*AND h.alcohol\_consumption IS NOT NULL*

*) AS health\_stats*

*ON 1=1*

*WHERE*

*h.life\_expectancy IS NOT NULL*

*AND h.air\_pollution IS NOT NULL*

*AND h.alcohol\_consumption IS NOT NULL;*

**Findings**

1. Life Expectancy:
   1. Countries in the Low Life Expectancy category typically lack access to quality healthcare and infrastructure.
   2. Countries with High Life Expectancy benefit from advanced healthcare systems and better living standards.
2. Air Pollution:
   1. Red and Purple zones (with air pollution levels >96) are concentrated in regions with industrial activities and lax environmental regulations.
   2. Green zones (<25) are associated with countries implementing sustainable energy practices and strict air quality laws.
3. Alcohol Consumption:
   1. Countries with High Alcohol Consumption may face public health challenges, including alcohol dependency and related health conditions.
   2. Low Alcohol Consumption often reflects cultural, religious, or regulatory factors.

**Implications**

1. Public Health: Addressing low life expectancy requires investment in healthcare infrastructure, especially in low-income regions. Alcohol consumption data highlights the need for awareness campaigns to reduce health risks.
2. Environmental Sustainability: High air pollution in certain regions underscores the need for stricter environmental policies and sustainable industrial practices. Countries with Green Zone air quality demonstrate the effectiveness of clean energy initiatives.

**Policy Recommendations**

1. Healthcare Investments:
   1. Focus on improving healthcare facilities and accessibility in regions with low life expectancy.
2. Environmental Regulations:
   1. Implement stricter emissions controls in countries within the Red and Purple zones.
3. Public Awareness Campaigns:
   1. Promote awareness about the risks of high alcohol consumption and encourage healthy lifestyles.

## CHALLENGES WHILE CREATING DATABASE

#### **Splitting and Insertion**

#### To ensure efficient and reliable data transfer, a staging table was employed alongside a Python script designed to streamline the migration process. The Python script executed optimized SQL transfer queries, which facilitated seamless movement of data from external sources into the main database. The staging table acted as a buffer to identify and handle data anomalies, ensuring data integrity in the main database while maintaining high transfer speeds and minimizing downtime during updates.

#### **Use of Derived Attributes**

A Generated Column was introduced to automate the computation and assignment of population density attribute values. This implementation dynamically calculated attribute values based on columns within the table, significantly reducing the need for manual intervention. By leveraging this approach, the system achieved efficient real-time updates with minimal overhead, as the derived values were computed on-the-fly during query execution. This setup was particularly advantageous in maintaining consistency across data while enhancing performance, as it eliminated the need for additional triggers or external scripts for updates.

#### **Handling Derived Attributes Dependent on Another Entity**

To manage derived attributes dependent on multiple related entities, a database view was created. The view joined Country and economy tables and dynamically calculated derived attribute GDP per capita during runtime, providing a more efficient and cleaner solution compared to alternatives like triggers or stored generated columns. By using a view, the system ensured consistency, maintainability, and reduced the complexity of updates, as changes in the underlying entities were automatically reflected in the derived attributes. Furthermore, the use of views simplified query design, improved data integrity, and enhanced the flexibility of the database structure without introducing additional storage overhead.

### ETHICS AND PRIVACY CONCERNS

In today's data-driven world, safeguarding privacy while generating valuable insights is paramount. This project takes a proactive approach to address ethical considerations and privacy protections, ensuring responsible use of data at the national level.

#### **Country-Level Aggregation**

The dataset utilized in this project is aggregated at the national level, significantly mitigating individual-level privacy risks. By focusing on broader trends and patterns across countries rather than personal data, the analysis eliminates any potential for identifying individuals. This ensures that all insights are derived without compromising individual privacy, fostering ethical data use while enabling comprehensive cross-country analysis of social and economic landscapes.

#### **Policy-Driven Analysis**

The analysis prioritizes actionable insights aimed at enhancing social and economic outcomes at the national level. Key metrics such as GDP per capita and labor force participation rates are analyzed to guide policymakers in addressing disparities and identifying areas needing support. The emphasis on generating constructive recommendations ensures that the analysis is not only ethical but also contributes positively to global development efforts.

#### **Privacy Protections**

1. **Exclusion of Personal Data**: The dataset does not include any personally identifiable information (PII), ensuring that privacy is never compromised.
2. **Focus on Aggregated Trends**: Insights are derived solely from aggregated data, further eliminating any risks associated with individual-level analysis.
3. **Transparency**: All methods and data sources are openly documented, ensuring accountability in the data handling process.

This approach ensures that the insights generated from the analysis are ethical, actionable, and aligned with the principles of data privacy and responsible usage.

**CONCLUSION**

This project explored key socioeconomic and health-related indicators across countries to uncover critical patterns and trends. By analyzing data on labor force participation, unemployment rates, literacy levels, GDP, and health metrics, we identified actionable insights that inform potential policy implementations.

The findings highlighted significant correlations between economic, social, and health indicators, such as the relationship between literacy rates and unemployment, the impact of GDP per capita on social well-being, and the categorization of countries based on air pollution and life expectancy. These analyses underscored disparities across nations and provided a foundation for targeted interventions aimed at fostering economic growth, improving public health, and enhancing environmental sustainability.

The project contributes to a broader understanding of global challenges, offering a data-driven basis for policymakers to address unemployment, improve education systems, support public health initiatives, and promote sustainable practices. By identifying vulnerable populations and areas requiring attention, this work supports the development of effective strategies for equitable and sustainable development.