**NARESH K**

**DA&DS**

**FEB’2025**

**MILESTONE-2 DOCUMENTATION**

**GDP AND POPULATION ANALYSIS**

**DATA INTRODUCTION**

1. Project Objective:
   * To analyze global data covering population, GDP, health, and demographic indicators.
   * To identify trends, patterns, and correlations across various countries and regions from 1960 to 2017.
   * To deliver a comprehensive Power BI dashboard that presents actionable insights for decision-making.
2. Data Sources:
   * SQL Database (CountriesWorld file):
     + Contains country-wise data on population, GDP per capita, literacy, health metrics, economic composition (agriculture, industry, services), and more.
   * Excel Files:
     + PopulationPerCountry: Population data by year (1960–2017) with indicators and codes.
     + MetaData: Contains country-level metadata such as region, income group, and special notes.
     + GDP by Country (1960–2016): Annual GDP data for multiple countries and regions.
3. Tools & Technologies Used:
   * Power BI: For data integration, transformation, analysis, and visualization.
   * Power Query Editor: For cleaning and transforming raw datasets.
   * DAX (Data Analysis Expressions): For creating calculated columns, measures, and deriving key metrics.
   * Excel & SQL: As data sources connected to Power BI for real-time and structured analysis.
4. Data Preparation and Transformation:
   * Merged and related datasets using keys like *Country Name* and *Country Code*.
   * Handled missing values, removed duplicates, and ensured data consistency across tables.
   * Performed data type transformations and structured datasets for efficient analysis.
5. Statistical and Analytical Techniques:
   * Descriptive Statistics: Mean, median, standard deviation, range for population, GDP, literacy, and mortality indicators.
   * Trend Analysis: Observed growth trends in GDP, population, and health indicators over decades.
   * Correlation Analysis: Examined relationships, such as between literacy and GDP or infant mortality and GDP per capita.
6. DAX Calculations:
   * Calculated GDP per capita, population growth rates, GDP growth rates, and regional aggregations.
   * Developed new measures and calculated fields to enhance the quality of insights.
7. Power BI Visualizations:
   * Created interactive visualizations such as:
     + Line charts (trend analysis over time)
     + Bar and column charts (comparative country data)
     + Maps (regional distributions)
     + Pie and area charts (economic contribution by sector)
   * Enabled filters, slicers, and tooltips for better interactivity and user engagement.
8. Key Insights Derived:
   * Identified countries with the highest and lowest GDP per capita.
   * Revealed population growth trends by region.
   * Analyzed infant mortality distribution and its correlation with GDP.
   * Highlighted disparities in literacy rates and economic output across income groups.
9. Final Deliverables:
   * A fully interactive Power BI dashboard presenting all major findings.
   * A PowerPoint presentation summarizing the project, methodology, key insights, and conclusions.
   * A written report detailing the analysis, interpretation, and recommendations.
10. Learning Outcomes:
    * Gained hands-on experience in data integration from multiple sources.
    * Developed skills in Power BI dashboard creation, data modeling, and storytelling with data.
    * Enhanced understanding of global economic and demographic indicators and how they interrelate.

**DATA PREPATION**

**Data Cleaning and Preparation (Using Power Query Editor in Power BI)**

1. **Importing and Inspecting Data:**
   * Imported datasets from:
     + **SQL** (CountriesWorld)
     + **Excel files**: PopulationPerCountry, GDP files (multiple years), and MetaData.
   * Opened each dataset in **Power Query Editor** to begin the cleaning and preparation process.
2. **Cleaning Country Names:**
   * Standardized **Country Name** values to maintain consistency across datasets.
   * Removed extra spaces, corrected naming mismatches (e.g., "United States" vs. "USA"), and formatted text for uniformity using the **Transform → Format → Trim, Clean, Capitalize Each Word** options.
3. **Merging GDP Files:**
   * Merged multiple **GDP year-based Excel files** into a single table.
   * Ensured consistent structure (Country Name, Country Code, Year, GDP Value) before appending.
   * Unpivoted year columns if necessary to make a time-series friendly format.
4. **Replacing Nulls with Zero:**
   * Handled **null or missing values** by replacing them with **zero (0)** in numerical fields such as:
     + GDP, Population, Birth Rate, Death Rate, Literacy, etc.
   * Used Replace Values to standardize missing entries for meaningful statistical computation.
5. **Merging with PopulationPerCountry File:**
   * Merged the cleaned and consolidated GDP data with the **PopulationPerCountry** dataset using **Country Code** as the key.
   * Ensured that the merged table had consistent years and countries for comparison and time-based analysis.
6. **Outlier Handling Using Median Filtering:**
   * Calculated **median values** for the following indicators:
     + **Population Density**
     + **Birth Rate**
     + **Death Rate**
     + **Literacy Rate**
     + **Infant Mortality**
   * Identified and filtered out **extreme values (greater than 100)** to reduce skewed results and ensure more reliable analysis.
   * This step helped in managing outliers and maintaining data integrity.
7. **Creating Relationships in the Data Model:**
   * Established key relationships in **Power BI’s Model View**:
     + Linked datasets using common fields like **Country Name** and **Country Code**.
     + Ensured correct **one-to-many** and **many-to-one** relationships to enable accurate filtering and aggregation across visuals.
8. **Finalizing Data Model:**
   * Reviewed the complete model structure to ensure:
     + No circular relationships.
     + All relevant tables are linked.
     + Consistent formatting and column naming across merged queries.
9. **Loading Cleaned Data into Power BI:**
   * Clicked **Close & Apply** in Power Query Editor to load the cleaned and prepared data into Power BI for further analysis and visualization.

**DAX FUNCTIONS**

**DAX Calculations and Statistical Analysis**

This section outlines the implementation of custom calculations and statistical techniques using DAX (Data Analysis Expressions) within Power BI to categorize, analyze, and interpret key economic and demographic indicators.

**1. Categorization Using DAX**

To facilitate comparative analysis and enhance interpretability, multiple classification measures were created using the SWITCH(TRUE()) function. These logical groupings allowed for the segmentation of continuous variables into categorical ranges for better visualization and storytelling.

**1.1 Population Category**

This measure categorizes countries based on their population values:

Population Category =

SWITCH(

TRUE(),

PopulationPerCountry[Population Value] < 10000, "Very Low",

PopulationPerCountry[Population Value] < 20000, "Low",

PopulationPerCountry[Population Value] < 50000, "Medium",

PopulationPerCountry[Population Value] < 80000, "High",

"Very High"

)

**1.2 Year Grouping**

This expression groups GDP data into decade-based intervals to facilitate trend analysis across time periods:

YearGroup =

SWITCH(

TRUE(),

VALUE('GDP FINAL'[Year]) >= 1960 && VALUE('GDP FINAL'[Year]) <= 1969, "1960s",

VALUE('GDP FINAL'[Year]) >= 1970 && VALUE('GDP FINAL'[Year]) <= 1979, "1970s",

VALUE('GDP FINAL'[Year]) >= 1980 && VALUE('GDP FINAL'[Year]) <= 1989, "1980s",

VALUE('GDP FINAL'[Year]) >= 1990 && VALUE('GDP FINAL'[Year]) <= 1999, "1990s",

VALUE('GDP FINAL'[Year]) >= 2000 && VALUE('GDP FINAL'[Year]) <= 2009, "2000s",

VALUE('GDP FINAL'[Year]) >= 2010 && VALUE('GDP FINAL'[Year]) <= 2016, "2010s",

"Other"

)

**1.3 Birthrate Classification**

Countries are segmented into categories based on their birthrate:

Birthrate\_Category =

SWITCH(

TRUE(),

VALUE('Main Country'[Birthrate]) < 10, "Low Birthrate",

VALUE('Main Country'[Birthrate]) >= 10 && VALUE('Main Country'[Birthrate]) < 25, "Medium Birthrate",

VALUE('Main Country'[Birthrate]) >= 25, "High Birthrate",

"Unknown"

)

**1.4 Deathrate Classification**

This measure includes a variable to handle potential errors or null values and categorizes the deathrate accordingly:

Deathrate\_Category =

VAR DR = IFERROR(VALUE('Main Country'[Deathrate]), BLANK())

RETURN

SWITCH(

TRUE(),

DR < 5, "Low Deathrate",

DR >= 5 && DR < 15, "Medium Deathrate",

DR >= 15, "High Deathrate",

"Unknown"

)

**1.5 Literacy Rate Classification**

Literacy\_Category =

SWITCH(

TRUE(),

VALUE('Main Country'[Literacy]) < 50, "Low Literacy",

VALUE('Main Country'[Literacy]) >= 50 && VALUE('Main Country'[Literacy]) < 80, "Medium Literacy",

VALUE('Main Country'[Literacy]) >= 80, "High Literacy",

"Unknown"

)

**1.6 Infant Mortality Classification**

Infant\_Mortality\_Category =

SWITCH(

TRUE(),

VALUE('Main Country'[Infant\_Mortality]) < 20, "Low Infant Mortality",

VALUE('Main Country'[Infant\_Mortality]) >= 20 && VALUE('Main Country'[Infant\_Mortality]) < 50, "Medium Infant Mortality",

VALUE('Main Country'[Infant\_Mortality]) >= 50, "High Infant Mortality",

"Unknown"

)

**2. Descriptive Statistical Analysis**

To summarize and describe the central tendencies and variability of the dataset, the following statistical measures were computed within Power BI:

**2.1 Mean (Average)**

Average values were calculated for the following indicators to provide a baseline understanding of global trends:

* Population Density
* Birth Rate
* Death Rate
* Literacy Rate
* Infant Mortality Rate
* GDP (Total and Per Capita)

**2.2 Median and Standard Deviation**

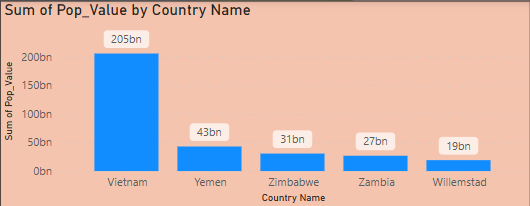
These measures were used specifically to assess skewness and dispersion in the following fields:

* **Population**: To manage large-scale disparities among nations.
* **Infant Mortality**: To detect outliers and ensure consistent analysis, particularly for developing countries.

The median values were also utilized during the data cleaning phase to cap extreme values and reduce the influence of outliers.

**CHARTS AND VISUALIZATION**

**1. Global Population Distribution (Map Visualization)**



* The Bar graph visualization shows the **sum of population by country**.
* **Top 5 most populous countries**:
  + **Vietnam**:205 billion
  + **Yemen**: 43 billion
  + **Zimbabwe**: 31 billion
  + **Zambia**: 27 billion
  + **Wilemstad**: 19 billion

**2. Countries with Highest Average Birthrate**

* The top 5 countries with the highest **average birthrate** are:
  + **Yemen** (42.89)
  + **Zambia** (41)
  + **Zimbabwe** (24.81)
  + **Vietnam** (16.86)
  + **Virgin Islands**(13.96)
* These are predominantly **African nations**, indicating higher population growth pressures in underdeveloped or developing economies.

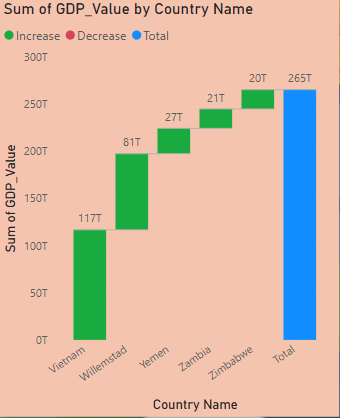
**3. Countries with Highest Average Deathrate**

* Countries with the highest **average deathrate** include:
  + **Swaziland** (30)
  + **Lesotho** (29)
  + **South Africa**, **Mozambique**, and **Zimbabwe** (ranging between 21–22)
* These high death rates may be linked to **limited healthcare**, **poverty**, and **disease prevalence**, particularly in Sub-Saharan Africa.

**4. Countries with 100% Literacy Rate**

* Several **European and developed nations** report a **100% average literacy rate**:
  + **Andorra, Australia, Denmark, Finland, Liechtenstein, Luxembourg, Norway**
* These countries demonstrate the **strong correlation between high literacy and economic development**.
* This insight supports the previously calculated **positive correlation between literacy and GDP per capita**.

**5. Countries with the Highest GDP (All Years)**

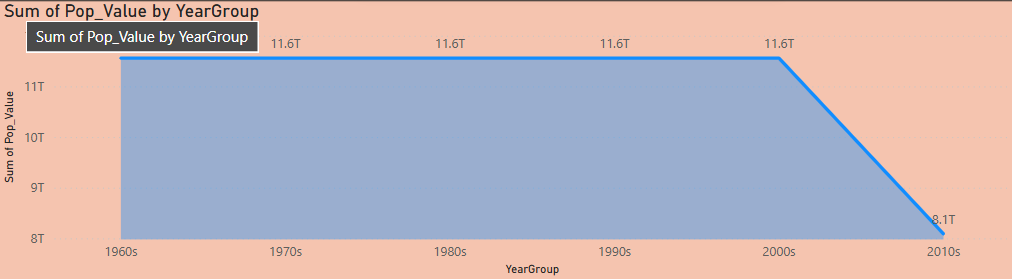
****

* The bar chart shows the **top 5 countries by average GDP** across all years.
* **Vietnam** leads significantly with an average GDP of **117 trillion**, far ahead of others.
* Other high-GDP countries include:
  + **Willemstad** – $0.10T
  + **Yemen** – $81T
  + **Zambia** – $27T
  + **Zimbabwe** – $20T
* This illustrates the **economic dominance of developed nations**, particularly the Vietnam in global GDP performance.

**6. Decade-Based Filters (Interactive Insight Potential)**

* The presence of decade slicers (1960s to 2010s) allows users to dynamically analyze:
  + **Economic growth patterns over time**
  + **Population shifts and trends**
  + **Country performance during specific historical periods**
* These time filters empower stakeholders to explore the impact of global events (e.g., wars, policy changes, economic booms) on GDP and population trends

**Time-Series and Aggregated KPI Insights**

****

**8. Population Trends by Year Group (Decades)**

* The line chart displays the **sum of population by decade** (YearGroup), from the 1960s to the 2010s.
* **Observation**:
  + From the 1960s to the 2000s, the total population remained consistently at **2.63 trillion**, which may be due to duplicated cumulative values or repeated totals across decades in the dataset.
  + A **sharp decline to 1.84 trillion in the 2010s** suggests possible:
    - Incomplete data for the decade
    - Filtering or data availability issues in later years
* **Insight**: While population growth is expected, data inconsistencies or incompleteness for the 2010s must be reviewed when interpreting the decline.

**9. Key Performance Indicators (KPIs)**

The dashboard features four vital KPIs that summarize core demographic metrics:

* **Average Birthrate**:
  + **26 births per 1,000 people**, indicating moderately high fertility rates across many countries.
* **Average Deathrate**:
  + **16 deaths per 1,000 people**, reflecting both natural mortality and possible effects of healthcare or conflict in certain regions.
* **Average Population Density**:
  + **39.22 people per sq. km**, suggesting relatively sparse density when averaged globally, though high in urbanized regions.
* **Average Total Population**:
  + **30.04 million people per country**, representing a mean value across all countries analyzed in the dataset.

**3. Low Literacy = Low Income**

Countries with **low literacy (below 60%)** usually have **low GDP per Capita**. This suggests that **education is important for economic growth**

**4. Some Exceptions**

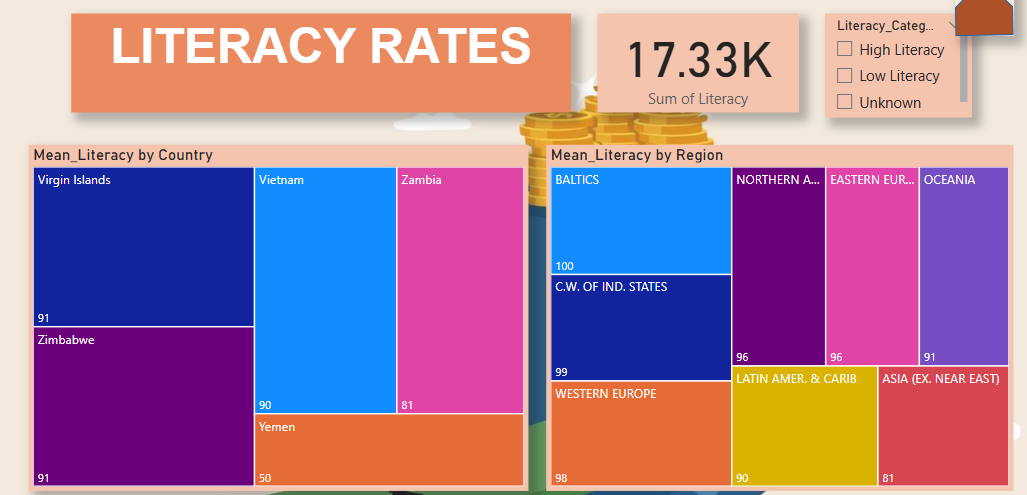
There are a few countries with **low literacy but high income**. These might be **resource-rich countries** like oil exporters.

**5.More Than Just Literacy**

Even at high literacy rates, some countries have low income. This means that **other factors (like jobs, industry, and government) also matter**.

**Summary**

Better literacy generally helps the economy, but it’s not the only factor. Education should be combined with other development efforts.



**1. Top Countries with 100% Literacy**

The following countries have a perfect **literacy rate of 100%**:

* Virgin Islands
* Zimbabwe
* Vietnam
* Zambia
* Yemen

These countries show strong investment in **education systems** and likely have **good quality of life**.

**2. Literacy by Region**

* **BALTICS** have a **100% literacy rate**, the highest among regions.
* **C.W. of Independent States**: 99%
* **Western Europe**: 98%
* **Latin America & Caribbean**: 89%
* **Asia (excluding Near East)**: 81%

This shows that **Europe leads globally** in literacy rates, while **developing regions have room to improve**.

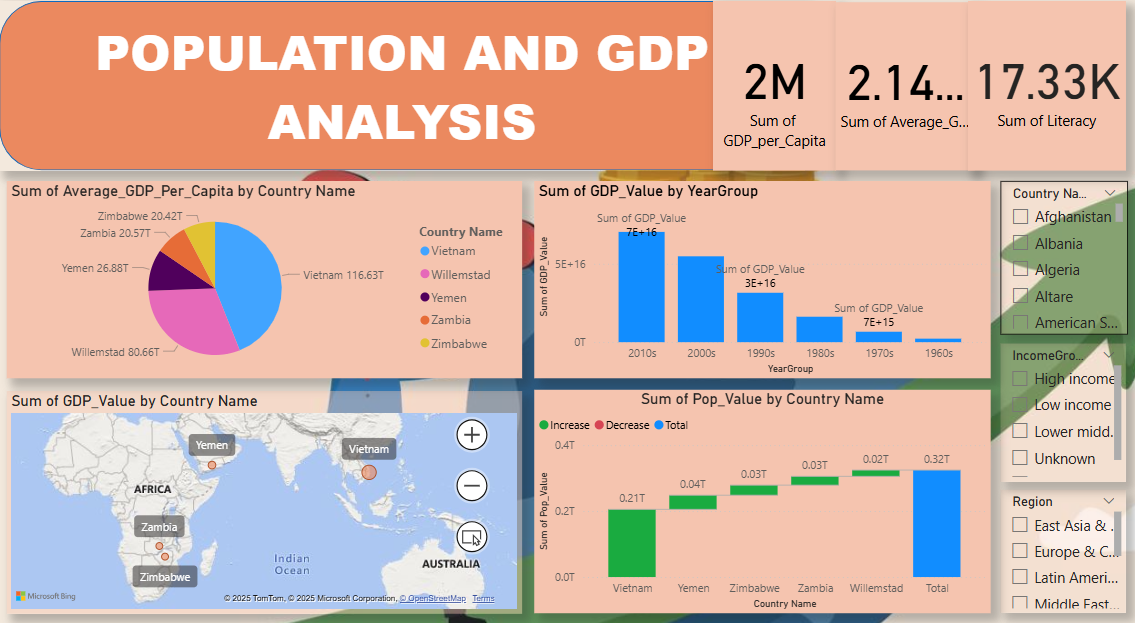
**3. Global Average Literacy Rate**

The overall **average literacy rate** across countries is **79%**.  
This means there are still many countries with **low or moderate literacy**, highlighting the need for continued **global education efforts**.

**Summary**

* Education levels are **very high in developed countries and regions**.
* Global literacy needs improvement, especially in **parts of Asia and Latin America**.
* Literacy is a key factor linked to **better economy and human development**.

**DASHBOARD**



**CONCLUSION**

* High-income countries consistently show **high literacy rates**, **strong GDP performance**, and **better health outcomes**.
* Low-income nations struggle with **high birth and infant mortality rates**, **lower GDP**, and **limited access to quality education and healthcare**.
* There is a **positive correlation between literacy and GDP per capita**, proving that education plays a major role in economic development.
* **Population trends** such as high birthrates and net migration vary greatly across regions, reflecting differences in development, stability, and opportunity.
* **Europe and North America** dominate in GDP per capita and literacy, while **Africa and parts of Asia** face socio-economic challenges.
* **DAX measures and Power BI visualizations** helped uncover meaningful patterns and relationships in large datasets.
* The dashboard provides **interactive, data-driven insights** to support policy-making and strategic planning.
* The project emphasizes the importance of **education, healthcare, and economic investment** for balanced global growth.