Global Analysis & Prediction of Life Expectancy Trends and Related Health Factors

Group- 2
Durga Bomma, Kay Meyers , Pallavi Vaswani, Saitejaswi Cherukupalli

Introduction

Analyzing data of time-period 2000-2015, this project investigates the factors influencing global life expectancy disparities, considering GDP, healthcare resources, and lifestyle behaviors.

Background

Examining the persistent global issue of child mortality, this research highlights the impactful links between socioeconomic factors, healthcare accessibility, and early childhood survival rates, emphasizing the need for a more nuanced understanding across diverse geographic and climatic settings.

Goals

Design a normalized relational database with tables connected through an Entity-Relationship Diagram to efficiently organize relevant project data.

Conduct statistical analysis guided by a research question to derive meaningful insights from the dataset.

Present visualization of data patterns from the statistical analysis to elucidate trends and aid understanding of the concepts explored.

Methodology

- Data Extraction & Data cleaning
- Database creation & Design
- Data Analysis
- Data visualization

Database creation

Server: 127.0.0.1:3308 » Database: group_2

Structure | SQL | Search | Query | Export | Import | Operations | Privileges | Routines | Events | Triggers | Designer

Filters

Containing the word:

Table	Action	Rows	Type	Collation	Size	Overhead
climate	Browse Structure Search Insert Empty Drop	2,864	InnoDB	utf8_general_ci	160.0 KiB	-
countries	Browse Structure Search Insert Empty Drop	179	InnoDB	utf8_general_ci	32.0 KiB	-
life_health	Browse Structure Search Insert Empty Drop	2,864	InnoDB	utf8_general_ci	352.0 KiB	-
year_modified	Browse Structure Search Insert Empty Drop	16	InnoDB	utf8_general_ci	16.0 KiB	-
4 tables	Sum	5,923	InnoDB	utf8mb4_general_ci	560.0 KiB	0 B

Check all   With selected:

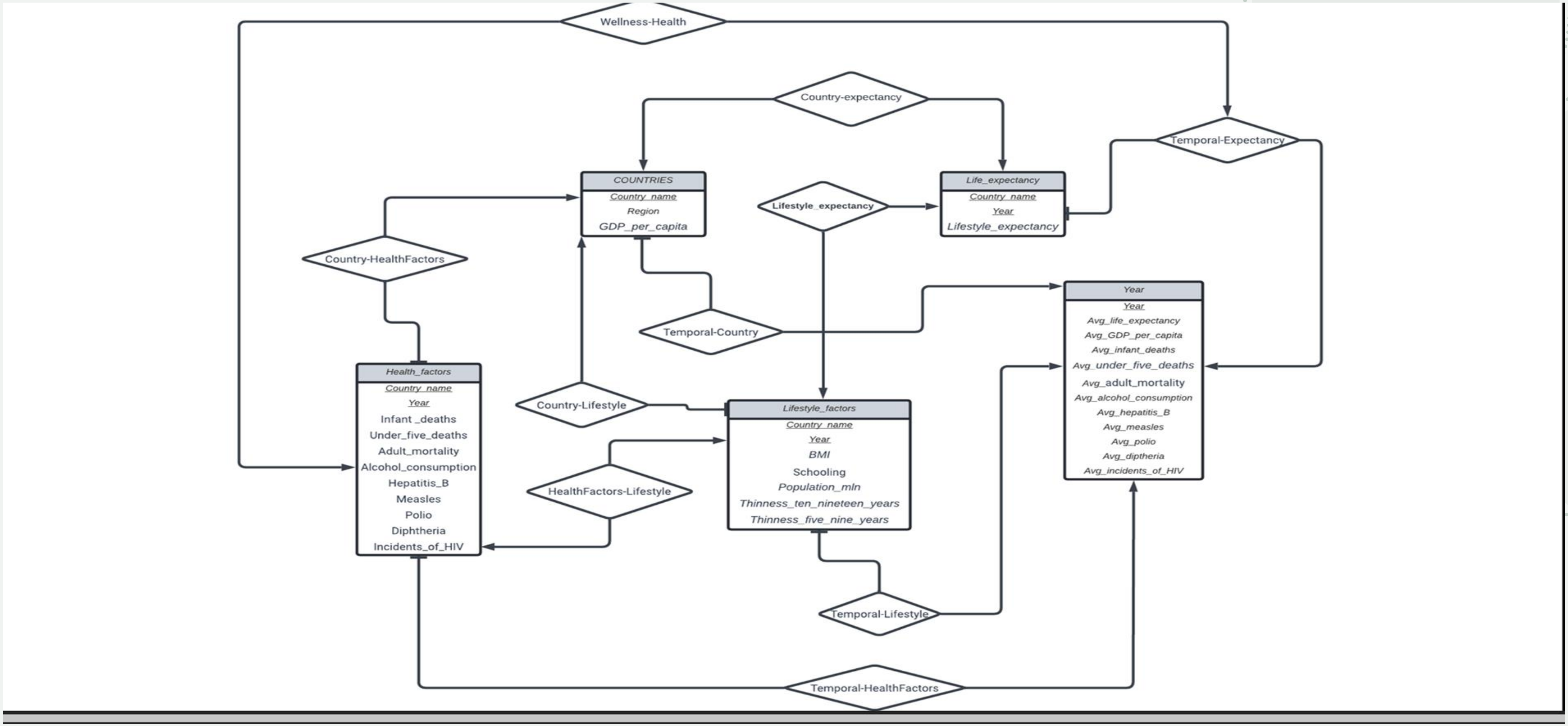
Print   Data dictionary

Create new table

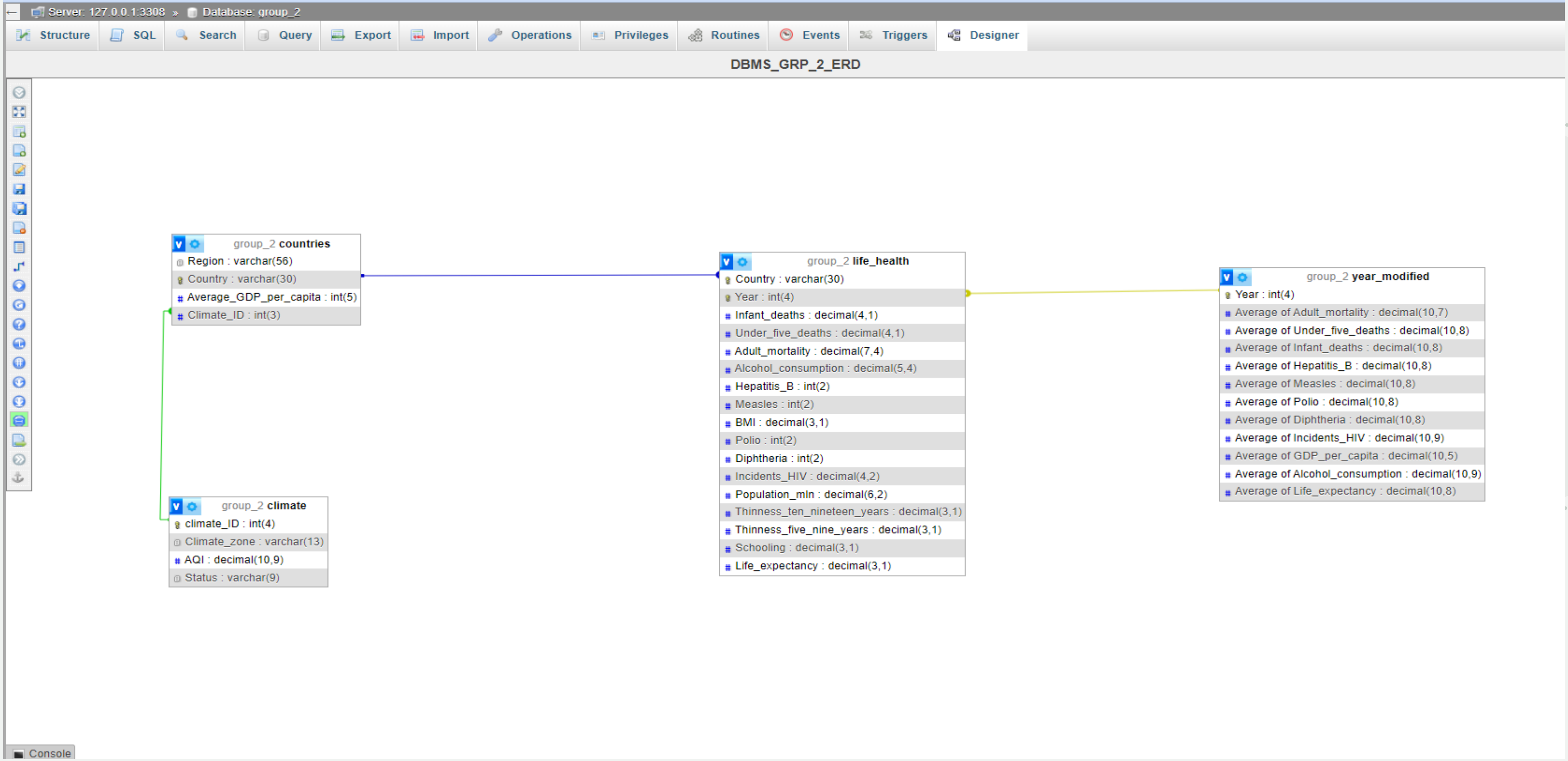
Table name   Number of columns   4   Create

Console

Preliminary ERD



Final ERD


Server: 127.0.0.1:3308 » Database: group_2
Structure
SQL
Search
Query
Export
Import
Operations
Privileges
Routines
Events
Triggers
Designer
DBMS_GRP_2_ERD
group_2 countries
Region : varchar(56)
Country : varchar(30)
Average_GDP_per_capita : int(5)
Climate_ID : int(3)
group_2 climate
climate_ID : int(4)
Climate_zone : varchar(13)
AQI : decimal(10,9)
Status : varchar(9)
group_2 life_health
Country : varchar(30)
Year : int(4)
Infant_deaths : decimal(4,1)
Under_five_deaths : decimal(4,1)
Adult_mortality : decimal(7,4)
Alcohol_consumption : decimal(5,4)
Hepatitis_B : int(2)
Measles : int(2)
BMI : decimal(3,1)
Polio : int(2)
Diphtheria : int(2)
Incidents_HIV : decimal(4,2)
Population_mln : decimal(6,2)
Thinness_ten_nineteen_years : decimal(3,1)
Thinness_five_nine_years : decimal(3,1)
Schooling : decimal(3,1)
Life_expectancy : decimal(3,1)
group_2 year_modified
Year : int(4)
Average of Adult_mortality : decimal(10,7)
Average of Under_five_deaths : decimal(10,8)
Average of Infant_deaths : decimal(10,8)
Average of Hepatitis_B : decimal(10,8)
Average of Measles : decimal(10,8)
Average of Polio : decimal(10,8)
Average of Diphtheria : decimal(10,8)
Average of Incidents_HIV : decimal(10,9)
Average of GDP_per_capita : decimal(10,5)
Average of Alcohol_consumption : decimal(10,9)
Average of Life_expectancy : decimal(10,8)
Console

Normalization- Part-1

Entity 1: Country

Attribute Type	Attribute	Description
Primary Key	Country_name	Unique identifier for each country
Non-key Attributes	Climate_ID	Identifier for the climate data
	Region	Geographical region of the country
	GDP_per_capita	Gross Domestic Product per capita
Normalization Status		Already in 3NF with atomic attributes, no repeating groups, partial dependencies, or transitive dependencies

Entity 2: Year

Attribute Type	Attribute	Description
Primary Key	Year	Unique identifier for each year
Non-key Attributes	Avg_life_expectancy	Average life expectancy for that year
	Avg_GDP_per_capita	Average Gross Domestic Product per capita
Normalization Status		Already in 3NF with atomic attributes, no repeating groups, partial dependencies, or transitive dependencies

Normalization- Part-2

Entity 3: Lifestyle_Health

Attribute Type	Attribute	Description
Composite Key	(Country_name, Year)	Combination of country and year as a unique identifier
Non-key Attributes	BMI	Body Mass Index
	Schooling	Education level or years of schooling
Normalization Status		Already in 3NF with atomic attributes, no repeating groups, partial dependencies, or transitive dependencies

Entity 4: Climate

Attribute Type	Attribute	Description
Primary Key	Climate_ID	Unique identifier for each climate record
Non-key Attribute	Climate_zone	Climate zone classification
	AQI	Air Quality Index
	Status	Status of the climate or environment
Normalization Status		Already in 3NF with atomic attributes, no repeating groups, partial dependencies, or transitive dependencies

Statistical Analysis

RESEARCH QUESTION

"How does the combination of economic status (GDP per capita) and healthcare access (represented by immunization rates for Hepatitis B, Polio, and Diphtheria) affect under-five mortality rates in various climate zones within developing countries over the last decade?"

STATISTICAL METHODS USED

- Shapiro-wilk- Normality test
- Spearman's Rank Correlation
- Kruskal-Wallis Test for Climate Zones
- Post-Hoc Analysis: Dunn's test

Shapiro-Wilk Normality Tests

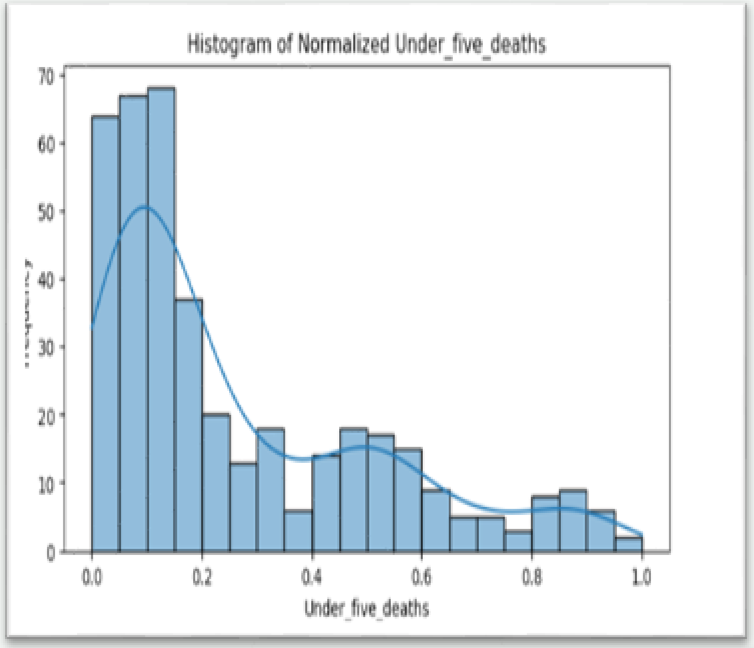
```
In [17]:
import pandas as pd
from scipy.stats import shapiro

# Assuming 'data' is your DataFrame with normalized data

# List of columns to test for normality
columns_to_test = ['Under_five_deaths', 'GDP_per_capita', 'Hepatitis_B', 'Polio', 'Diphtheria'

# Performing Shapiro-Wilk test on each column
for column in columns_to_test:
    stat, p = shapiro(data[column])
    print(f'Normality test for {column}: Statistics={stat:.3f}, p={p:.3f}')

    # Interpretation
    alpha = 0.05
    if p > alpha:
        print(f'  {column} looks Gaussian (fail to reject H0)')
    else:
        print(f'  {column} does not look Gaussian (reject H0)')
```

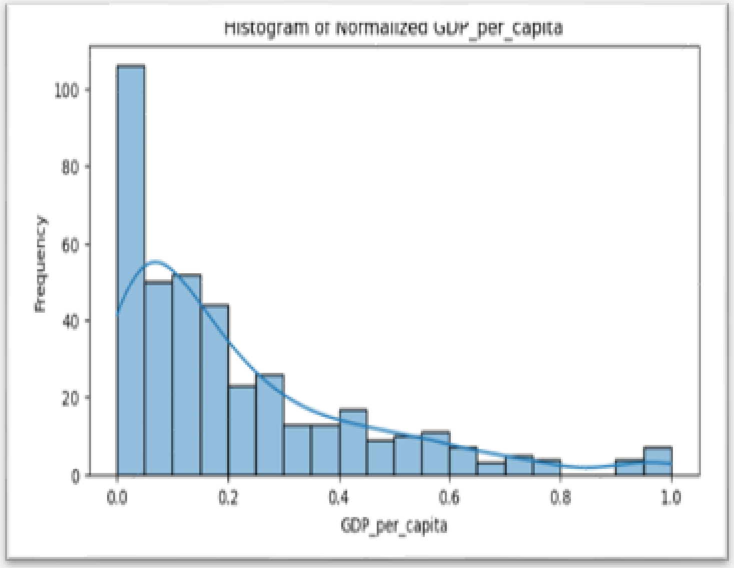
```
Normality test for Under_five_deaths: Statistics=0.851, p=0.000
  Under_five_deaths does not look Gaussian (reject H0)
Normality test for GDP_per_capita: Statistics=0.828, p=0.000
  GDP_per_capita does not look Gaussian (reject H0)
Normality test for Hepatitis_B: Statistics=0.866, p=0.000
  Hepatitis_B does not look Gaussian (reject H0)
Normality test for Polio: Statistics=0.851, p=0.000
  Polio does not look Gaussian (reject H0)
Normality test for Diphtheria: Statistics=0.856, p=0.000
  Diphtheria does not look Gaussian (reject H0)
```

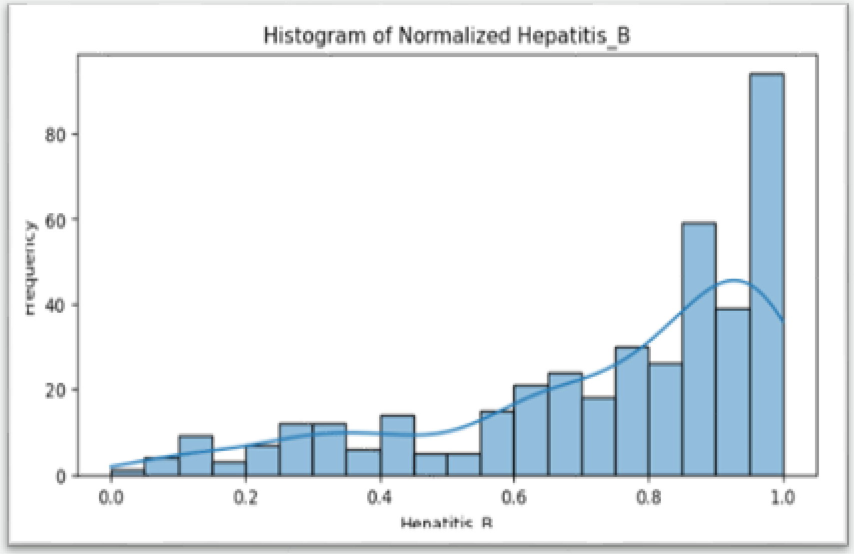
- The results from your Shapiro-Wilk normality tests indicate that the data in all the tested columns ('Under_five_deaths', 'GDP_per_capita', 'Hepatitis_B', 'Polio', and 'Diphtheria') do not follow a normal distribution (Gaussian distribution).

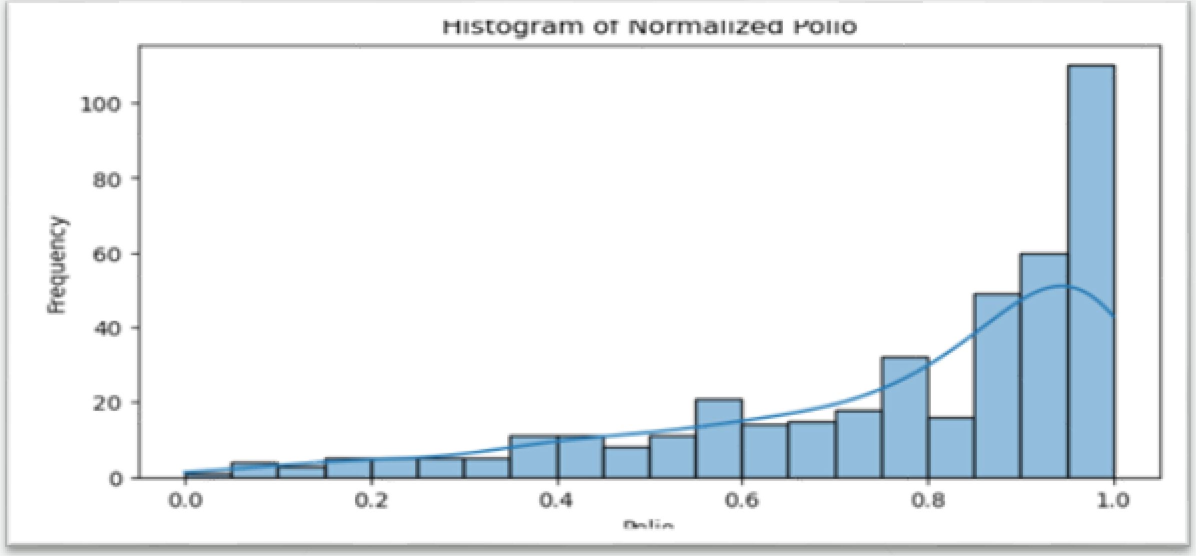
- This conclusion is drawn from the fact that the p-values for all these columns are very small (0.000), leading to the rejection of the null hypothesis that the data is normally distributed.

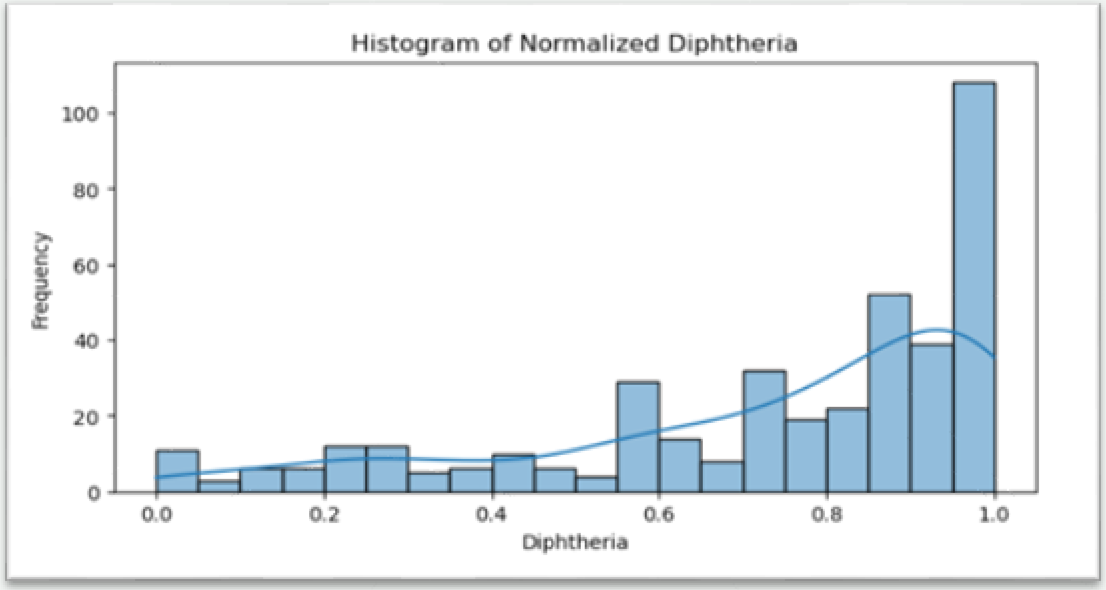
Histograms of Each Variable











Spearman's Rank Correlation

- To assess the relationship between continuous variables such as 'GDP_per_capita', 'Hepatitis_B', 'Polio', 'Diphtheria', and 'Under_five_deaths'.

```
import pandas as pd
from scipy.stats import spearmanr

# Assuming 'data' is your DataFrame
columns_to_correlate = ['GDP_per_capita', 'Hepatitis_B', 'Polio', 'Diphtheria', 'Under_five_deaths']

# Calculating Spearman's Rank Correlation
for col1 in columns_to_correlate:
    for col2 in columns_to_correlate:
        if col1 != col2:
            coef, p = spearmanr(data[col1], data[col2])
            print(f"Spearman correlation between {col1} and {col2}: Coefficient={coef:.3f}, P-value={p:.3f}")
```

```
Spearman correlation between GDP_per_capita and Hepatitis_B: Coefficient=0.340, P-value=0.000
Spearman correlation between GDP_per_capita and Polio: Coefficient=0.391, P-value=0.000
Spearman correlation between GDP_per_capita and Diphtheria: Coefficient=0.410, P-value=0.000
Spearman correlation between GDP_per_capita and Under_five_deaths: Coefficient=-0.823, P-value=0.000
Spearman correlation between Hepatitis_B and GDP_per_capita: Coefficient=0.340, P-value=0.000
Spearman correlation between Hepatitis_B and Polio: Coefficient=0.884, P-value=0.000
Spearman correlation between Hepatitis_B and Diphtheria: Coefficient=0.936, P-value=0.000
Spearman correlation between Hepatitis_B and Under_five_deaths: Coefficient=-0.427, P-value=0.000
Spearman correlation between Polio and GDP_per_capita: Coefficient=0.391, P-value=0.000
Spearman correlation between Polio and Hepatitis_B: Coefficient=0.884, P-value=0.000
Spearman correlation between Polio and Diphtheria: Coefficient=0.919, P-value=0.000
Spearman correlation between Polio and Under_five_deaths: Coefficient=-0.493, P-value=0.000
Spearman correlation between Diphtheria and GDP_per_capita: Coefficient=0.410, P-value=0.000
Spearman correlation between Diphtheria and Hepatitis_B: Coefficient=0.936, P-value=0.000
Spearman correlation between Diphtheria and Polio: Coefficient=0.919, P-value=0.000
Spearman correlation between Diphtheria and Under_five_deaths: Coefficient=-0.495, P-value=0.000
Spearman correlation between Under_five_deaths and GDP_per_capita: Coefficient=-0.823, P-value=0.000
Spearman correlation between Under_five_deaths and Hepatitis_B: Coefficient=-0.427, P-value=0.000
Spearman correlation between Under_five_deaths and Polio: Coefficient=-0.493, P-value=0.000
Spearman correlation between Under_five_deaths and Diphtheria: Coefficient=-0.495, P-value=0.000
```

Interpretation:

1. **Positive Correlations with GDP Per Capita**: There are positive correlations between GDP per capita and immunization rates (Hepatitis B, Polio, Diphtheria), suggesting that higher economic status is generally associated with better immunization coverage.

2. **Negative Correlations with Under-Five Deaths**: There are strong negative correlations between under-five mortality rates and both GDP per capita and immunization rates. This indicates that higher economic status and better immunization coverage are associated with lower under-five mortality rates.

Kruskal-Wallis Test for Climate Zones

To evaluate how under-five mortality rates vary across different climate zones, the Kruskal-Wallis test can be used. This test is the non-parametric version of ANOVA and is used when comparing more than two groups.

```
import pandas as pd
from scipy.stats import shapiro

# Assuming 'data' is your DataFrame with normalized data

# List of columns to test for normality
columns_to_test = ['Under_five_deaths', 'GDP_per_capita', 'Hepatitis_B', 'Polio', 'Diphtheria'

# Performing Shapiro-Wilk test on each column
for column in columns_to_test:
    stat, p = shapiro(data[column])
    print(f'Normality test for {column}: Statistics={stat:.3f}, p={p:.3f}')

    # Interpretation
    alpha = 0.05
    if p > alpha:
        print(f'  {column} looks Gaussian (fail to reject H0)')
    else:
        print(f'  {column} does not look Gaussian (reject H0)')
```

```
Normality test for Under_five_deaths: Statistics=0.851, p=0.000
  Under_five_deaths does not look Gaussian (reject H0)
Normality test for GDP_per_capita: Statistics=0.828, p=0.000
  GDP_per_capita does not look Gaussian (reject H0)
Normality test for Hepatitis_B: Statistics=0.866, p=0.000
  Hepatitis_B does not look Gaussian (reject H0)
Normality test for Polio: Statistics=0.851, p=0.000
  Polio does not look Gaussian (reject H0)
Normality test for Diphtheria: Statistics=0.856, p=0.000
  Diphtheria does not look Gaussian (reject H0)
```

Interpretation

Statistical Significance: The very low p-value (0.000) suggests that there are statistically significant differences in under-five mortality rates among the various climate zones in the dataset.

Post-Hoc Analysis: Dunn's test

Since the Kruskal-Wallis test indicates that there are differences but does not specify between which climate zones these differences occur, that is why conducting post-hoc tests.
Methods like the Dunn's test can be used to compare specific pairs of climate zones to identify where the significant differences lie.

```python
import pandas as pd
from scipy.stats import kruskal
from statsmodels.stats.multicomp import pairwise_tukeyhsd, MultiComparison

# Assuming 'data' is your DataFrame and has columns 'Climate_Zone' and 'Under_five_deaths'

# Conducting Kruskal-Wallis Test
climate_zones = data['Climate_Zone'].unique()
grouped_data = [data['Under_five_deaths'][data['Climate_Zone'] == zone] for zone in climate_zones]
stat, p = kruskal(*grouped_data)
print(f"Kruskal-Wallis Test: Statistics={stat:.3f}, p={p:.3f}")

# Conducting Dunn's Post-Hoc Test
mc = MultiComparison(data['Under_five_deaths'], data['Climate_Zone'])
result = mc.tukeyhsd()

print(result)
print(mc.groupsunique)
```

```
Kruskal-Wallis Test: Statistics=152.658, p=0.000
      Multiple Comparison of Means - Tukey HSD, FWER=0.05
==================================================================
    group1        group2     meandiff p-adj   lower    upper  reject
------------------------------------------------------------------
       Desert       Diverse  -0.0171 0.9999 -0.2331  0.1989  False
       Desert           Dry   0.0833 0.5763 -0.0627  0.2294  False
       Desert Mediterranean  -0.2193 0.0059 -0.3967  -0.042   True
       Desert     Temperate  -0.1319 0.0397 -0.2602 -0.0037   True
       Desert      Tropical   0.1032 0.1073 -0.0118  0.2182  False
      Diverse           Dry   0.1004 0.7584 -0.1131  0.3139  False
      Diverse Mediterranean  -0.2022 0.1408 -0.4383  0.0338  False
      Diverse     Temperate  -0.1148 0.5795 -0.3166   0.087  False
      Diverse      Tropical   0.1203 0.4805 -0.0733  0.3139  False
          Dry Mediterranean  -0.3027    0.0  -0.477 -0.1283   True
          Dry     Temperate  -0.2152    0.0 -0.3393 -0.0912   True
          Dry      Tropical   0.0198 0.9956 -0.0904  0.1301  False
Mediterranean     Temperate   0.0874 0.6205 -0.0723  0.2472  False
Mediterranean      Tropical   0.3225    0.0  0.1732  0.4718   True
    Temperate      Tropical   0.2351    0.0  0.1498  0.3204   True
------------------------------------------------------------------
['Desert' 'Diverse' 'Dry' 'Mediterranean' 'Temperate' 'Tropical']
```

Tropical Zone > Desert Zone > Dry Zone > Temperate Zone > Mediterranean Zone

Analysis Inference

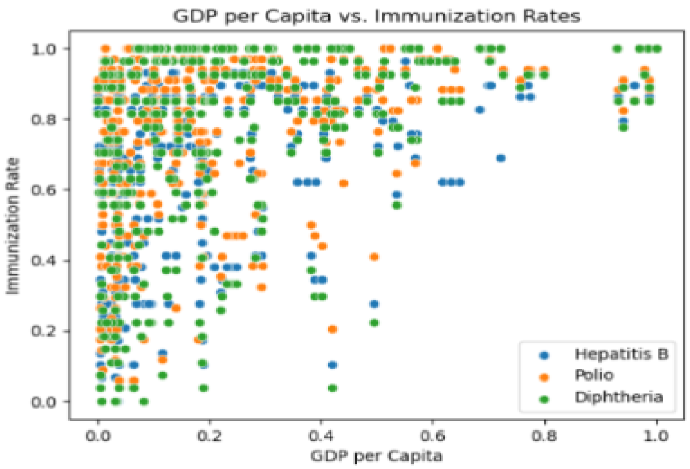
1. GDP per Capita and Healthcare Access: Countries with higher GDP per capita usually have better healthcare access, including higher immunization rates for Hepatitis B, Polio, and Diphtheria.

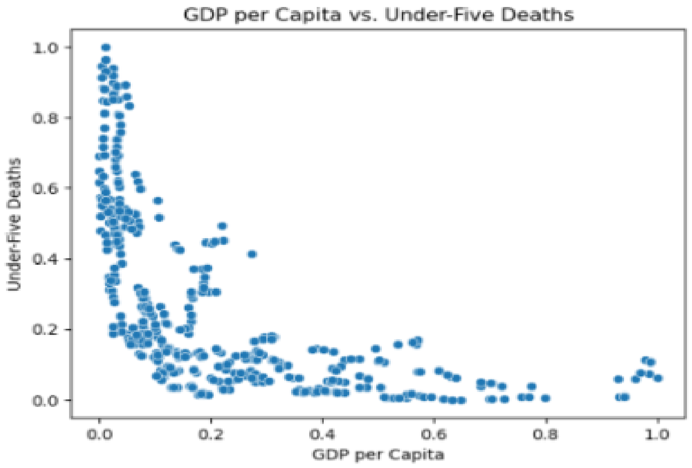
2. Vaccinations and Under-Five Deaths: When more children get vaccinated, the number of under-five deaths decreases. This shows that better healthcare access helps in reducing child mortality.

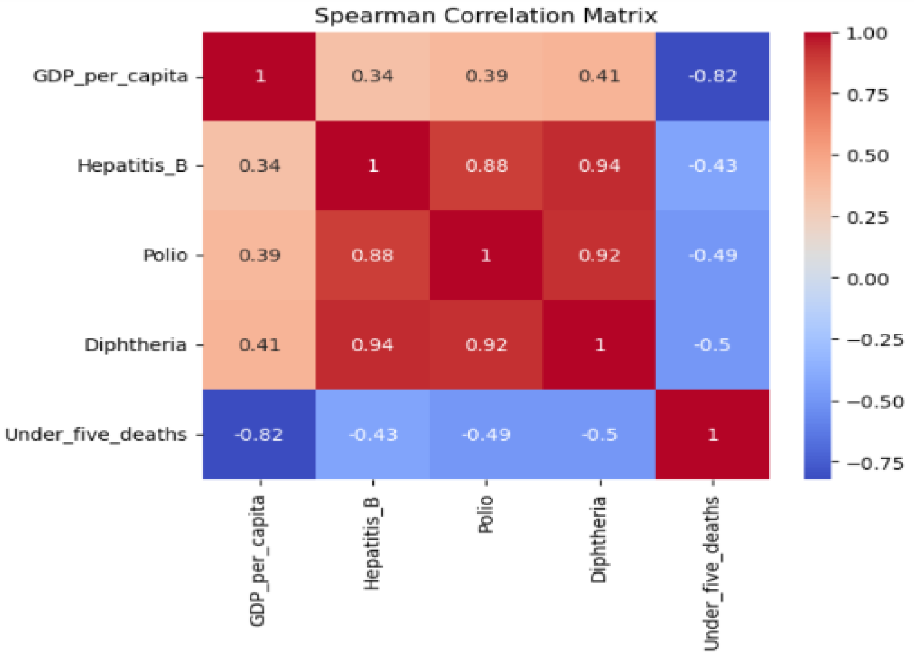
3. Different Climates, Different Results: The impact of GDP per capita and healthcare access on under-five deaths varies depending on the climate zone. For instance, in Tropical climates, under-five deaths are higher compared to Mediterranean or Temperate climates.

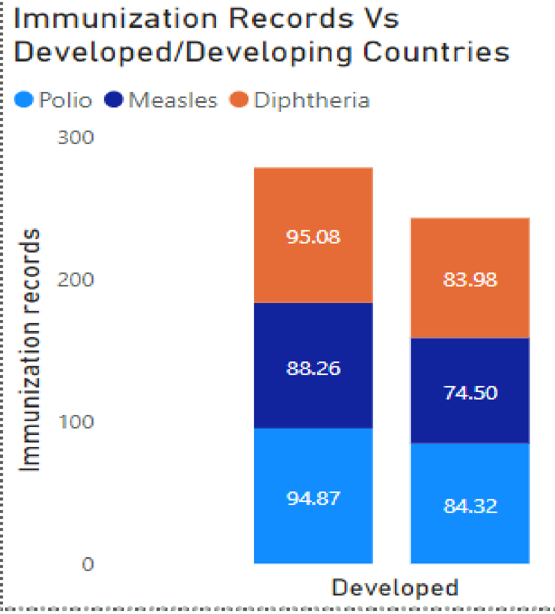
4. Climate's Influence: The study found that the climate zone plays a role in the number of under-five deaths in those areas.

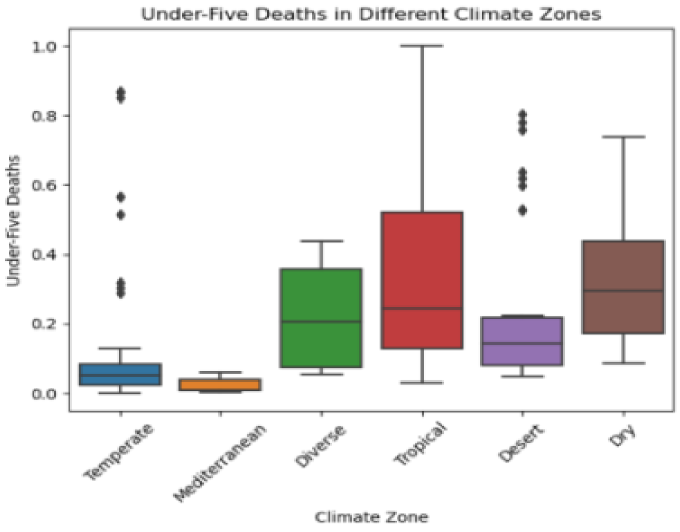
Dashboard based on Research Question

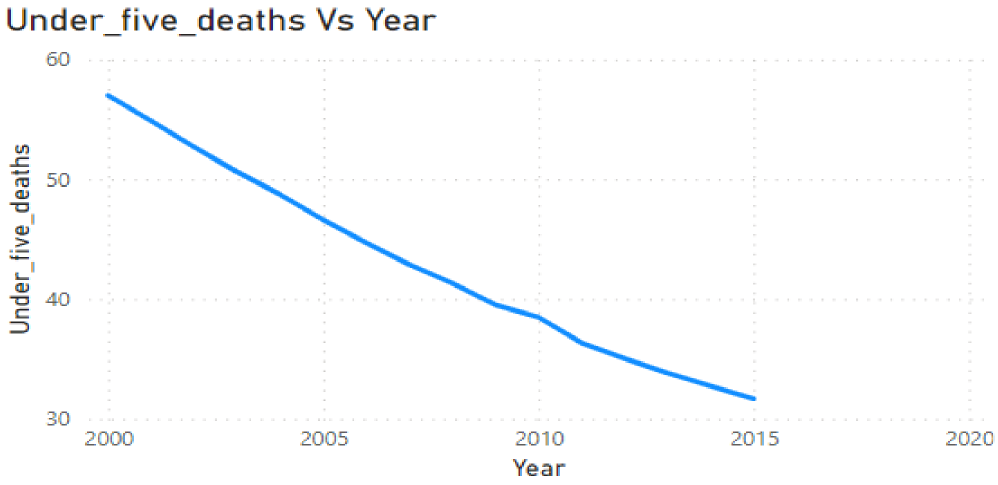




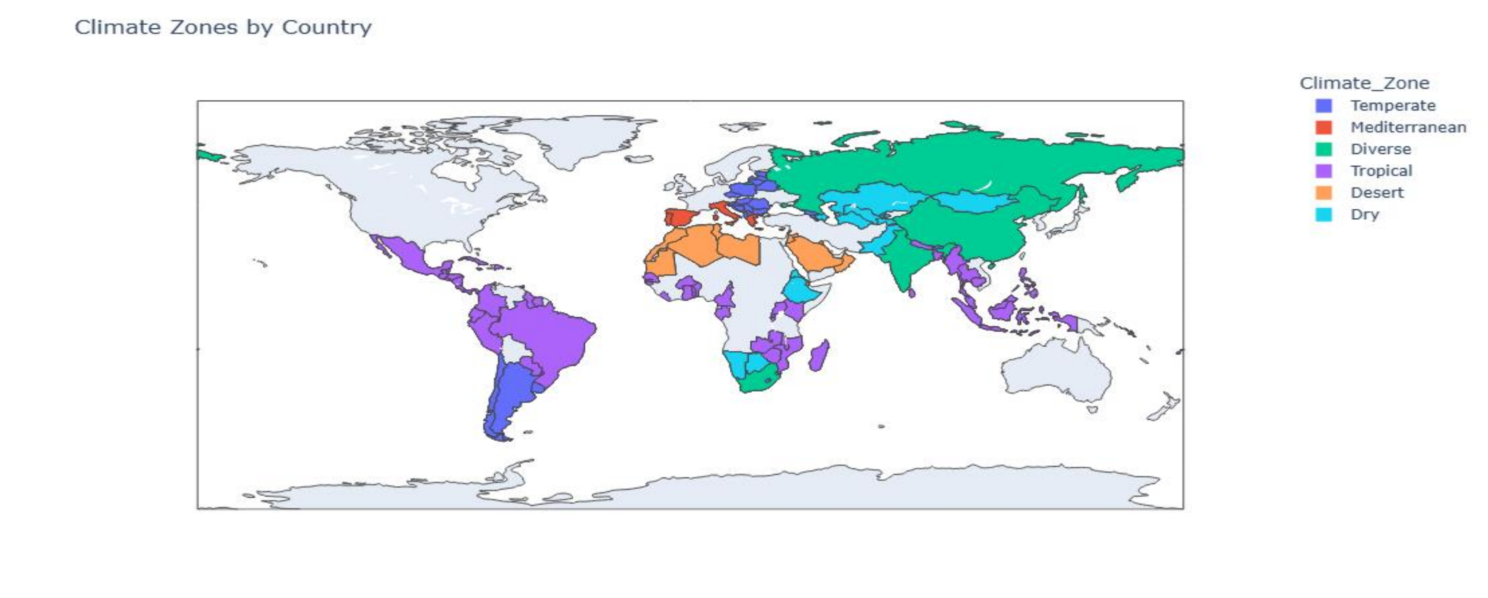


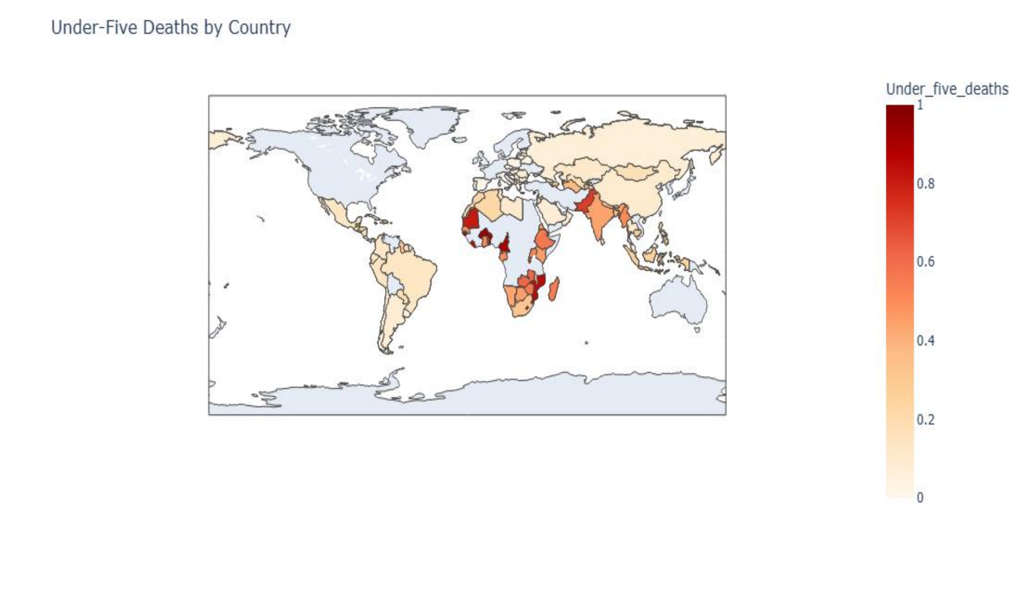


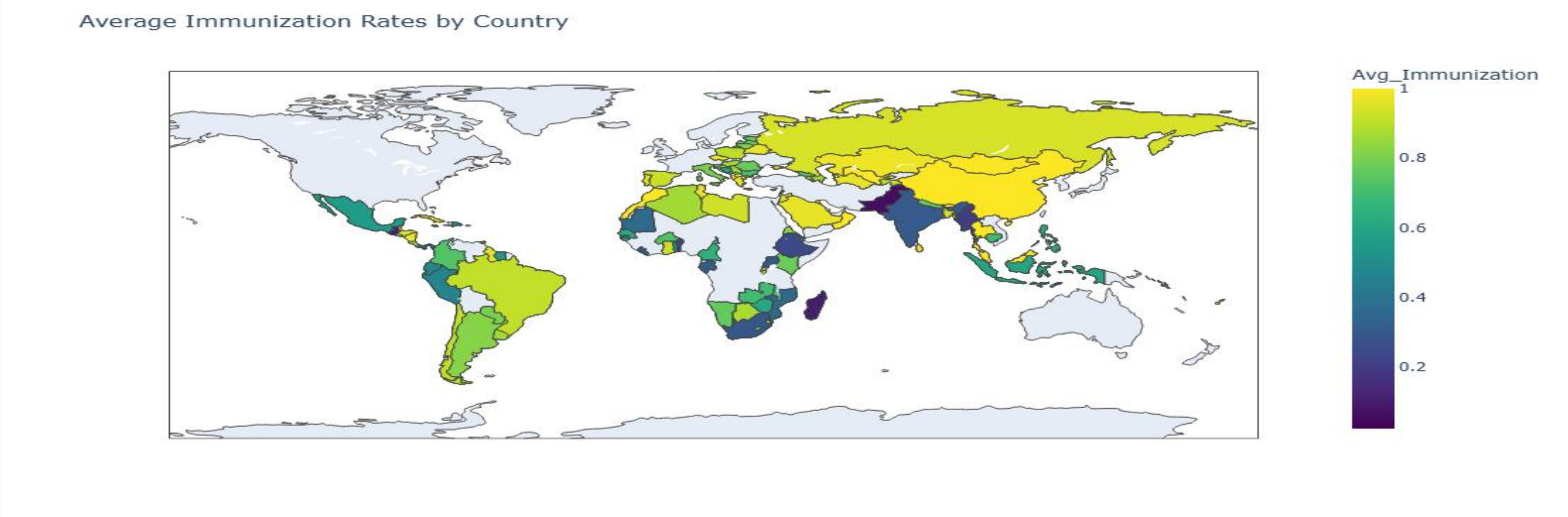




Chloropleth Maps

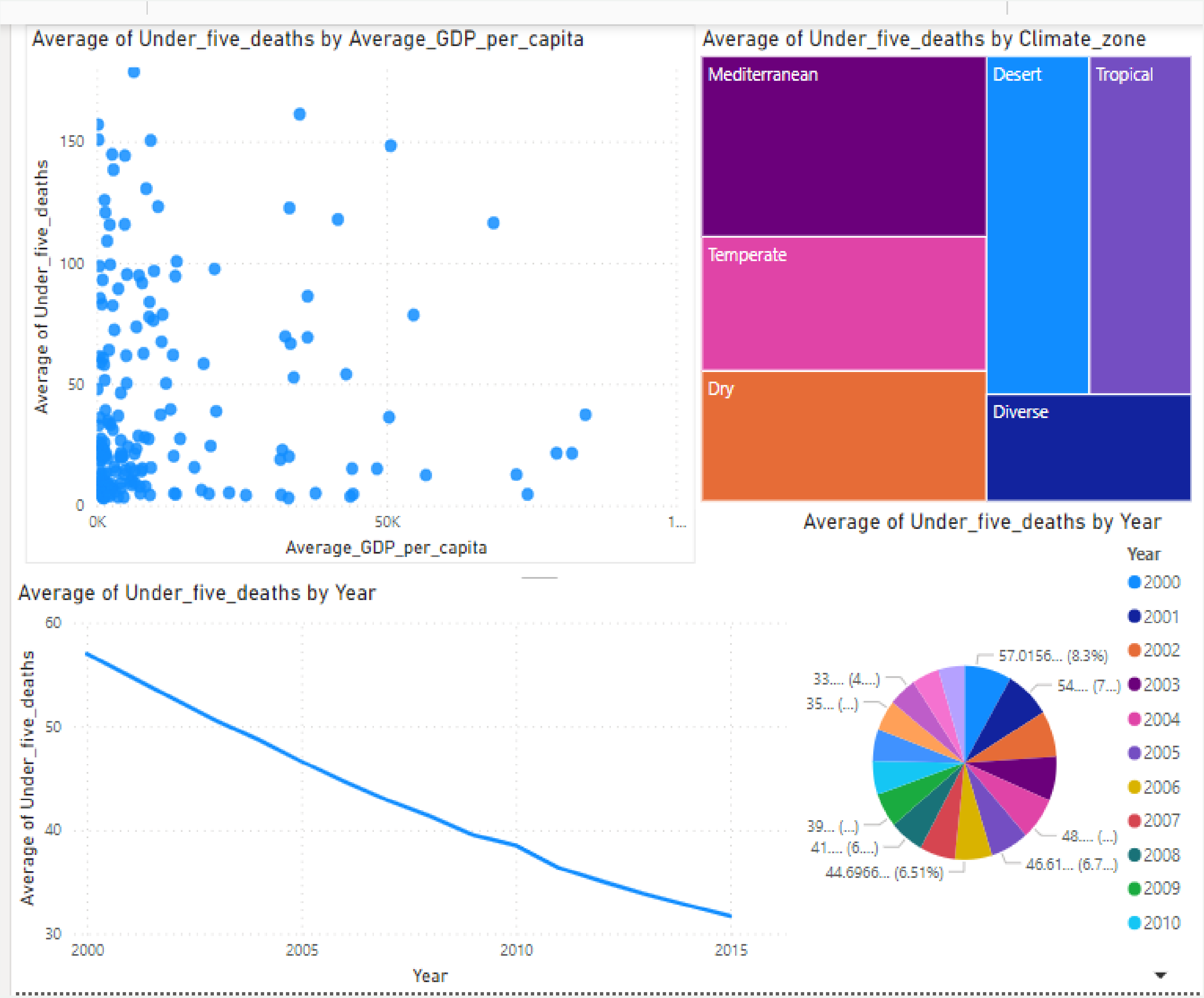






Link to Statistical Analysis:
[Life expectancy project]

Visualizations and Queries



SQL QUERIES FOR VISUALIZATIONS

Avg_Under_5_deaths by Avg_GDP_percapita

SELECT
AVG(Under_5_deaths)
AS Avg_Under_5_deaths,
AVG(GDP_percapita) AS
Avg_GDP_percapita
FROM Year;

Avg_Under_5_deaths by Year

SELECT Year, AVG(Under_5_deaths) AS
Avg_Under_5_deaths
FROM Year GROUP BY Year ORDER
BY Year;

Database Design Influence on Data Analytics Success

Data Integrity & Consistency: Our adherence to the Third Normal Form has been pivotal in ensuring accuracy and reliability in our analytics.

Efficient Data Access: The structured entities like Country and Climate facilitate quick data retrieval, crucial for our timely analytics.

Integration and Relationship Management: The way we've modeled relationships between entities such as Country, Climate, and Lifestyle_Health enables comprehensive, multi-dimensional analysis. This interconnectedness is key for drawing deeper insights from our data.

Proposed Enhancements for Future Analytics

Expand Dataset

- Widen data scope by incorporating more countries, additional correlate variables, and longer timespan.

Enrich Analysis

- Employ more advanced statistical and predictive modeling like multivariate regression, neural networks, time series forecasting, and outlier analysis.

Deeper Regional Analysis

Interactive Visualization

Challenges

While working on the project, the one difficulty we have faced is connecting the SQL server with Python and the Visualization Tool.

Team Responsibilities

Durga Bomma	Database normalization, Database documentation, Database design using SQL, Presentation, Report writing.
Kay Meyers	Report writing.
Pallavi Vaswani	E-R diagram design, Database normalization, Data analysis and visualizations with python, Presentation, Report writing.
Saitejaswi Cherukupalli	Database normalization, Database documentation, Visualizations using Power BI, Presentation, Report writing.

THANK YOU