Exploring the Relationship Between Environmental Quality, Economic Prosperity, and Life Expectancy in 2010 to 2015

Group 21
Yicong Li
Jono Liang
Miaoyuan Wang
Nicole Zhu
Business School, University of Auckland
BUSINFO 702 Information Management
August 2025

Executive Summary

This project explores the connections between economic development, environmental quality, and public health across countries. Its goal is to understand how nations can achieve economic growth while maintaining clean air and healthy populations.

The study uses international datasets covering GDP per capita, CO_2 emissions, PM2.5 air pollution levels, life expectancy, and country codes (See Appendix A for data source details). These indicators allow us to compare economic performance, environmental impacts, and health outcomes across continents and countries.

Key insights show that higher pollution levels are generally linked to lower life expectancy, though outcomes vary across regions depending on policies and healthcare capacity. Economic growth does not always mean higher pollution, as some countries achieve both high GDP and good air quality through effective regulations. CO_2 emissions are closely tied to development but can be mitigated with cleaner technologies. Finally, the CHEE Index highlights which countries balance health, environment, and economic performance most effectively.

These findings emphasize that evaluating progress requires more than GDP alone. Integrating environmental and health metrics can guide policies toward sustainable and equitable growth, helping nations improve wellbeing while reducing environmental impact.

Research Question

Based on the chosen datasets, we have come up with five questions.

Question 1

What is the relationship between air pollution (PM2.5) and life expectancy across different continents, and how do countries within each continent differ in their PM2.5 exposure?

Exposure to fine particulate matter (PM2.5) is a major global health concern, linked to millions of premature deaths each year. Studies have shown that higher PM2.5 levels significantly reduce lifespan, making it crucial to assess these impacts across countries and continents (Burnett et al., 2018). Understanding the relationship between pollution levels and life expectancy can highlight which regions are most vulnerable and inform policies to protect public health.

Question 2

Within each continent, how does the relationship between GDP per capita and PM2.5 levels evolve over the years (2010 - 2015)?

Economic growth often brings industrialization and increased emissions, but it can also provide resources for pollution control. The Environmental Kuznets Curve suggests that as countries become wealthier, pollution may first increase and then decrease (Grossman & Krueger, 1995). Investigating this relationship over multiple years can reveal whether higher GDP consistently improves or worsens air quality in different regions.

Question 3

How are countries distributed across continents under different PM2.5 band levels?

Categorizing countries by PM2.5 bands (e.g., Good, Moderate, Unhealthy) provides a clearer picture of the global air quality landscape. This question helps identify which regions face the highest environmental risks and how widespread exposure to unhealthy air is. Understanding this distribution is critical for targeted interventions and prioritizing environmental policies.

Question 4

How do CO_2 emissions evolve over time across continents, and what is their relationship with economic development (GDP) between 2010 and 2015?

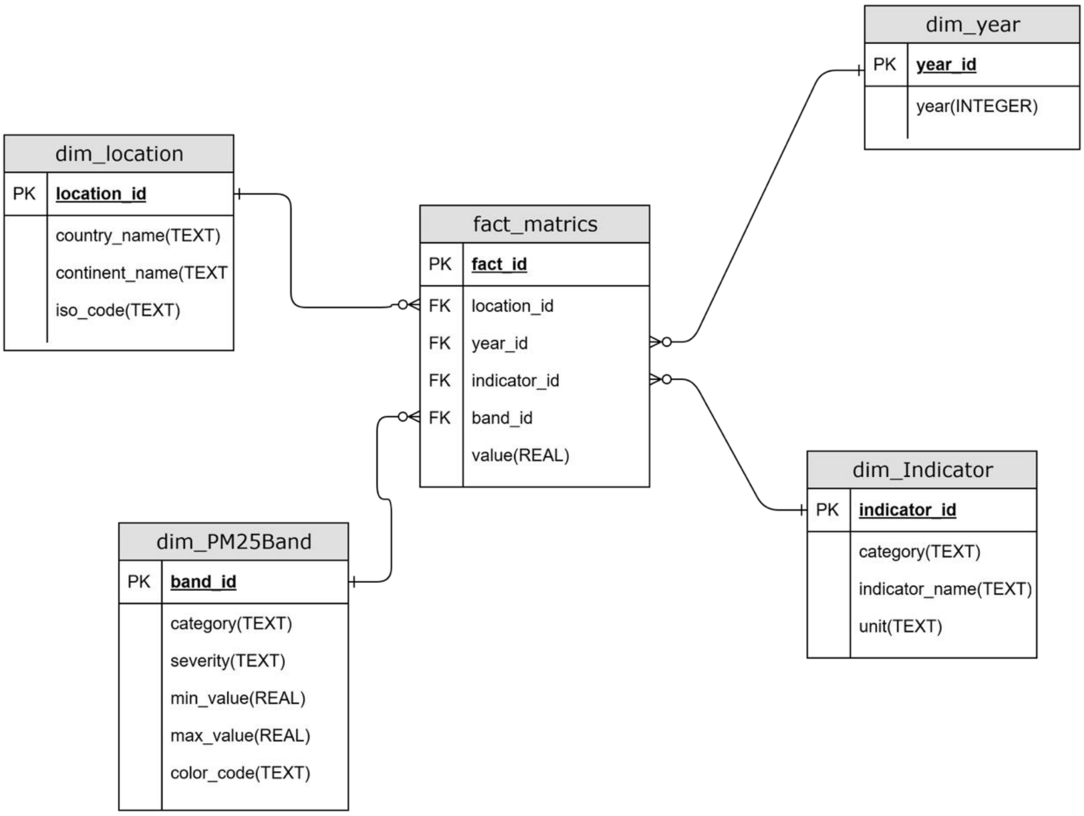
CO_2 emissions reflect industrial and economic activity, but they also contribute to environmental degradation. Some countries have managed to decouple economic growth from rising emissions, while others have not (Wang & Su, 2020). This question investigates how emissions patterns change over time and whether higher GDP necessarily leads to higher emissions, offering insights for sustainable development strategies.

Question 5

How do countries and continents perform on the Composite Health–Economy–Environment Index based on formula we designed?

The CHEE Index combines life expectancy, PM2.5 pollution levels, GDP per capita, and CO_2 emissions into a single measure to capture the trade-offs between health, environmental quality, and economic development. This question evaluates which countries achieve the best balance and which lag behind, providing insight into sustainable and equitable growth.

Star Schema Design



The star schema is designed to support analytical queries on environmental metrics, specifically focusing on air quality measured by PM2.5 values. At the center of the schema is the **Fact_Metrics** table, which stores the quantitative measurements of PM2.5 for different locations and time periods, along with associated categorical attributes such as bands (e.g., "Good," "Moderate," "Unhealthy") that indicate the severity of pollution levels.

Fact Table: Fact_Metrics

The **Fact_Metrics** table is the central fact table in the star schema. Each record in this table represents a measurement of PM2.5 at a particular location and time.

The fact table links to multiple dimension tables, allowing for detailed analysis and aggregation across different perspectives.

Dimension Tables

1. **Dim_Location**:

   This table provides geographic context to each measurement, containing attributes such as country and continent. It allows us to analyze PM2.5 levels across locations or drill down from continent to country-level summaries. The hierarchical structure supports queries such as GDP per capita trend or PM2.5 trend by location.

2. **Dim_Time**:
   This table allows time-based aggregations and trend analysis, such as CO_2 emissions over time.
3. **Dim_Indicator**:
   This table defines indicator measurements for each category, such as category Environment measured by PM2.5 and CO_2 emissions, category Economy measured by GDP per capita, and category Health measured by life expectancy.
4. **Dim_Band**:
   This table provides descriptive labels and thresholds for PM2.5 readings, enabling us to categorize numeric measurements into meaningful health-impact bands. For example, a PM2.5 value of 20 might map to the "Moderate" band. This facilitates easier reporting and visualization of air quality levels.

ELT Implementation in SQLite

Extract

The raw datasets were collected from multiple CSV files, including PM2.5 air pollution data, GDP per capita, CO_2 emissions, life expectancy, and a country codes table. These were imported into SQLite as staging tables (stg_pm25, stg_gdp, stg_life, stg_country). Before loading, inconsistent column names were standardized and unnecessary characters were cleaned.

Code:

```
--ALTER TABLE "stg_pm25_long" RENAME TO stg_pm25;
--ALTER TABLE "GDPpercapita" RENAME TO stg_gdp;
--ALTER TABLE "LifeExpectancyData" RENAME TO stg_life;
--ALTER TABLE "country_codes" RENAME TO stg_country;
--ALTER TABLE "lifeexpectancy" RENAME TO stg_life2;
```

Load

The CSV data was then loaded into SQLite using the .import command. Some old or duplicate tables were dropped to avoid conflicts.

Code:

```
DROP TABLE IF EXISTS stg_pm25;
DROP TABLE IF EXISTS "Life Expectancy Data";
```

We ensured consistency by activating referential integrity:

Code:

```
PRAGMA foreign_keys = ON;
```

Transform

To enable multidimensional analysis, the data was transformed into a **star schema** with several dimension tables and one fact table.

1. Time Dimension

Code:

```
CREATE TABLE Dim_Year (
  year_id INTEGER PRIMARY KEY,
  year    INTEGER NOT NULL
);
```

This table assigns a unique key to each year (only include 2010 - 2015).

```
--INSERT INTO Dim_Year (year_id, year) VALUES
```

```
--(1, 2010),
--(2, 2011),
--(3, 2012),
--(4, 2013),
--(5, 2014),
--(6, 2015)
```

2. Location Dimension

```
CREATE TABLE Dim_Location (
 location_id   INTEGER PRIMARY KEY,
 country_name   TEXT NOT NULL,
 continent_name TEXT,
 iso_code      TEXT
);
```

This table links countries to their continent and ISO code.

```
--INSERT INTO Dim_Location (country_name, continent_name, iso_code)
--SELECT DISTINCT
-- c1, -- Country Name
-- c3, -- Region
-- c2 -- Country Code (ISO)
--FROM stg_life2
--WHERE c1 IS NOT NULL AND TRIM(c1) <> '';
```

3. Indicator Dimension

```
CREATE TABLE Dim_Indicator (
 indicator_id  INTEGER PRIMARY KEY,
 category      TEXT NOT NULL,
 indicator_name TEXT NOT NULL,
 unit          TEXT NOT NULL
);
```

This table was created to classify and describe all indicators used in the analysis, including their category (Environment, Economy, or Health), name (PM2.5, CO_2 emissions, GDP per capita, Life expectancy), and measurement unit, so that each fact record can be consistently linked to its context.

```
INSERT INTO Dim_Indicator (indicator_id, category, indicator_name, unit)
VALUES
(1,'Environment','PM2.5','µg/m³'),
(2,'Environment','CO2_emissions','kiloton'),
(3,'Economy','GDP_per_capita','USD'),
(4,'Health','Life_expectancy','years');
```

4. PM2.5 Band Dimension

Code:

```
CREATE TABLE dim_PM25Band (
  band_id    INTEGER PRIMARY KEY AUTOINCREMENT,
  category   TEXT      NOT NULL,
  severity   TEXT,
  min_value  REAL      NOT NULL,
  max_value  REAL      NOT NULL,
  color_code TEXT      NOT NULL
);
```

The dim_PM25Band table defines PM2.5 bands, each with a unique band_id, category, severity, min_value, max_value, and color_code. It maps numeric PM2.5 values to air quality levels, facilitating easy interpretation of pollution severity.

```
INSERT INTO dim_PM25Band (category, severity, min_value, max_value,
color_code) VALUES
('Good',     NULL,     0,   75,  '#00FF00'),
('Polluted', 'Light',  75,  115, '#FFFF00'),
('Polluted', 'Moderate', 115, 150, '#FFA500'),
('Polluted', 'Heavy',  150, 250, '#FF4500'),
('Polluted', 'Severe', 250, 500, '#FF0000');
```

5. Fact Table

Code:

```
CREATE TABLE Fact_Metrics (
 fact_id      INTEGER PRIMARY KEY,
 location_id  INTEGER NOT NULL,
```

```
 year_id      INTEGER NOT NULL,
 indicator_id  INTEGER NOT NULL,
 value        REAL,
 band_id      INTEGER,
 FOREIGN KEY(location_id)  REFERENCES Dim_Location(location_id),
 FOREIGN KEY(year_id)      REFERENCES Dim_Year(year_id),
 FOREIGN KEY(indicator_id) REFERENCES Dim_Indicator(indicator_id),
 FOREIGN KEY(band_id)      REFERENCES Dim_PM25Band(band_id)
);
```

The **Fact_Metrics** table stores all numeric values, linked by foreign keys to the dimension tables:

Populating the Fact Table

PM2. 5 Values with Bands

```
INSERT INTO Fact_Metrics (location_id, year_id, indicator_id, value, band_id)
SELECT dL.location_id, dY.year_id, dI.indicator_id,
       round(p.c4,2), b.band_id
FROM stg_pm25 p
JOIN Dim_Location dL ON p.c2 = dL.iso_code OR UPPER(TRIM(p.c1)) =
UPPER(TRIM(dL.country_name))
JOIN Dim_Year dY ON dY.year = p.c3
JOIN Dim_Indicator dI ON dI.indicator_name = 'PM2.5'
LEFT JOIN Dim_PM25Band b ON p.c4 >= b.min_value AND p.c4 < b.max_value
WHERE p.c3 BETWEEN 2010 AND 2015
  AND p.c4 IS NOT NULL;
```

This SQL statement inserts PM2.5 data into the fact table by linking each record with its country, year, indicator, and pollution band level, while filtering valid values between 2010 and 2015.

CO_2 Emissions

```
INSERT INTO Fact_Metrics (location_id, year_id, indicator_id, value)
SELECT dl.location_id, dy.year_id, 2, round(p.c8,2)
FROM stg_life2 p
JOIN Dim_Location dl ON TRIM(UPPER(p.c1)) = TRIM(UPPER(dl.country_name))
JOIN Dim_Year dy ON dy.year = p.c5
WHERE p.c8 IS NOT NULL;
```

This SQL statement loads CO_2 emissions data into the fact table by matching each record with its country and year, ensuring only non-null values are stored.

GDP per capita

```
INSERT INTO Fact_Metrics (location_id, year_id, indicator_id, value)
SELECT dl.location_id, dy.year_id, 3, round(p.c17,2)
FROM stg_life p
JOIN Dim_Location dl ON UPPER(TRIM(p.c1)) = UPPER(TRIM(dl.country_name))
JOIN Dim_Year dy ON dy.year = p.c2
WHERE p.c17 IS NOT NULL;
```

This SQL statement inserts GDP per capita data into the fact table by linking each record with its country and year, while filtering out null values.

Life Expectancy

```
INSERT INTO Fact_Metrics (location_id, year_id, indicator_id, value)
SELECT dl.location_id, dy.year_id, 4, p.c4
FROM stg_life p
JOIN Dim_Location dl ON UPPER(TRIM(p.c1)) = UPPER(TRIM(dl.country_name))
JOIN Dim_Year dy ON dy.year = p.c2
WHERE p.c4 IS NOT NULL;
```

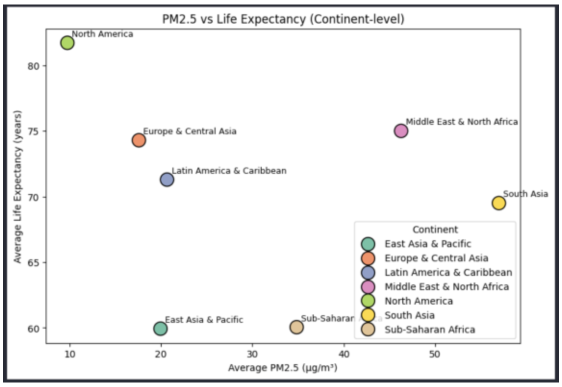
This SQL statement inserts life expectancy data into the fact table by joining country and year information, ensuring only valid records are included.

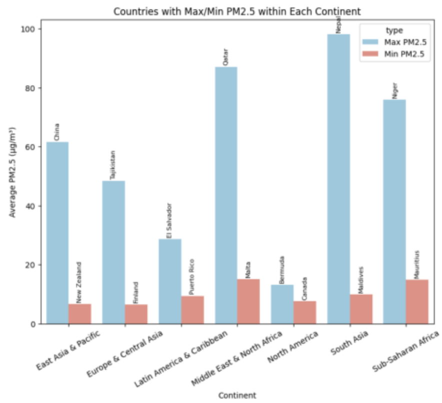
SQL for Business Analytics

Please see Appendix B for all SQL statements used for analysing below questions.

Research Question 1

What is the relationship between air pollution (PM2.5) and life expectancy across different continents, and how do countries within each continent differ in their PM2.5 exposure?





Interpretation

The first chart shows a negative relationship between average PM2.5 levels and life expectancy across continents. North America stands out with very low pollution (below 10 µg/m³) and the highest life expectancy (above 82 years), while South Asia has the highest pollution (around 60 µg/m³) and lower life expectancy (about 69 years). Sub-Saharan Africa is an exception, where life expectancy remains low despite only moderate PM2.5 levels, suggesting other influences such as healthcare and poverty.

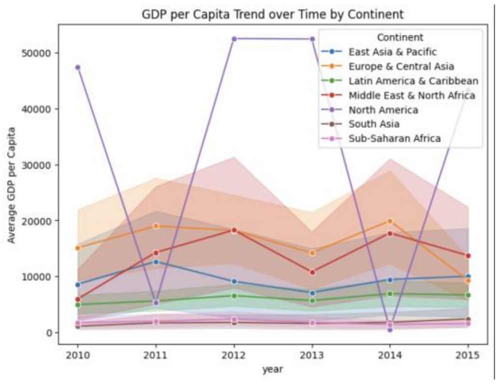
The second chart highlights large contrasts within continents. In East Asia & Pacific, China records very high PM2.5 while New Zealand remains among the cleanest. In Europe & Central Asia, Tajikistan experiences heavy pollution whereas Finland enjoys low levels. South Asia shows the widest gap, with Nepal reaching nearly 100 µg/m³ compared to the Maldives' much cleaner air.

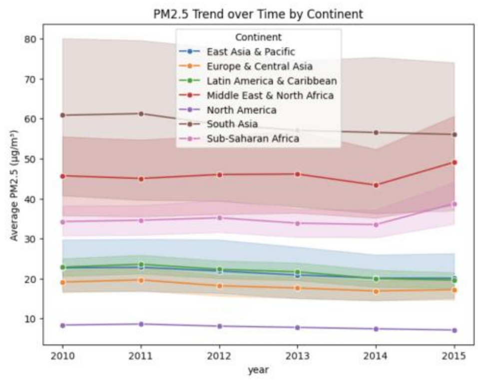
Taken together, the two charts indicate that while higher PM2.5 is generally linked to shorter life expectancy, outcomes vary significantly across and within continents.

National policies, economic development, and healthcare capacity play an important role in mediating the health impacts of air pollution.

Research Question 2

Within each continent, how does the relationship between GDP per capita and PM2.5 levels evolve over the years (2010 - 2015) ?





Interpretation

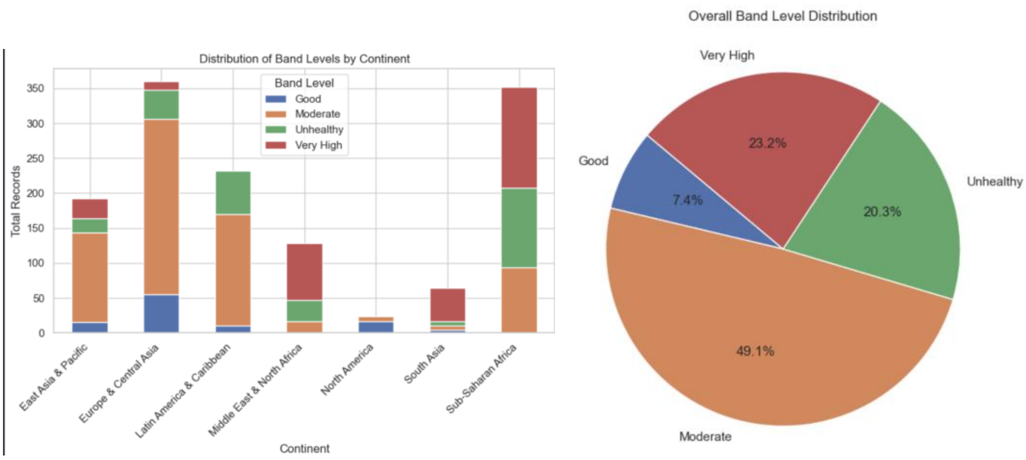
The first plot shows that GDP per capita varies widely between continents, with North America consistently having the highest values, despite some visible fluctuations across years. Europe & Central Asia and the Middle East & North Africa also maintain relatively higher GDP per capita compared to other regions, while South Asia and Sub-Saharan Africa remain at the lowest end. This demonstrates the persistent global inequality in economic development.

The second plot highlights the trends in PM2.5 levels during the same period. South Asia consistently exhibits the highest average PM2.5 concentration, with values above 55 µg/m³, followed by the Middle East & North Africa. In contrast, North America maintains the lowest PM2.5 levels, remaining below 10 µg/m³ throughout the period. Most regions show relatively stable PM2.5 patterns, though there is a slight decline in East Asia & Pacific, Europe & Central Asia, and Latin America & Caribbean, suggesting some improvements in air quality over time.

When compared together, the two graphs suggest that higher GDP per capita does not necessarily correlate with higher pollution levels. For example, North America combines high GDP per capita with low PM2.5, while South Asia shows the opposite pattern, with low GDP per capita but very high pollution levels. This indicates that economic growth can coincide with both improvement and deterioration in air quality, depending on regional policies, industrial structures, and environmental regulations.

Research Question 3

How are countries distributed across continents under different PM2.5 band levels?



Interpretation

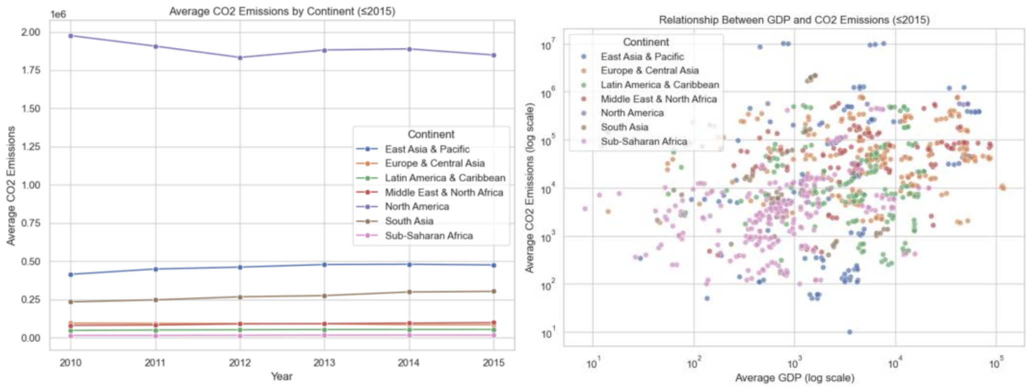
The stacked bar chart on the left shows how countries across different continents are distributed under various PM2.5 band levels. **Europe & Central Asia** and **Latin America & Caribbean** are dominated by the “Moderate” category, while **Sub-Saharan Africa** has a large share in the “Very High” and “Unhealthy” categories, highlighting severe air quality challenges. In contrast, **North America** has relatively fewer records overall, but most fall into “Good” or “Moderate,” indicating better air quality. These differences underscore how exposure to air pollution varies significantly by region.

The pie chart on the right provides a global overview of PM2.5 band levels. Nearly half of all records fall into the “Moderate” category (49.1%), followed by “Very High” (23.2%) and “Unhealthy” (20.3%), while only **7.4%** of records fall under “Good.” This global distribution emphasizes that clean air remains limited, and the majority of the world’s population experiences air quality worse than “Good.”

Overall, the pie chart highlights the **overall imbalance in global air quality**, while the stacked bar chart illustrates **where the disparities are most pronounced geographically**.

Research Question 4

How do CO_2 emissions evolve over time across continents, and what is their relationship with economic development (GDP) between 2010 and 2015?



Interpretation

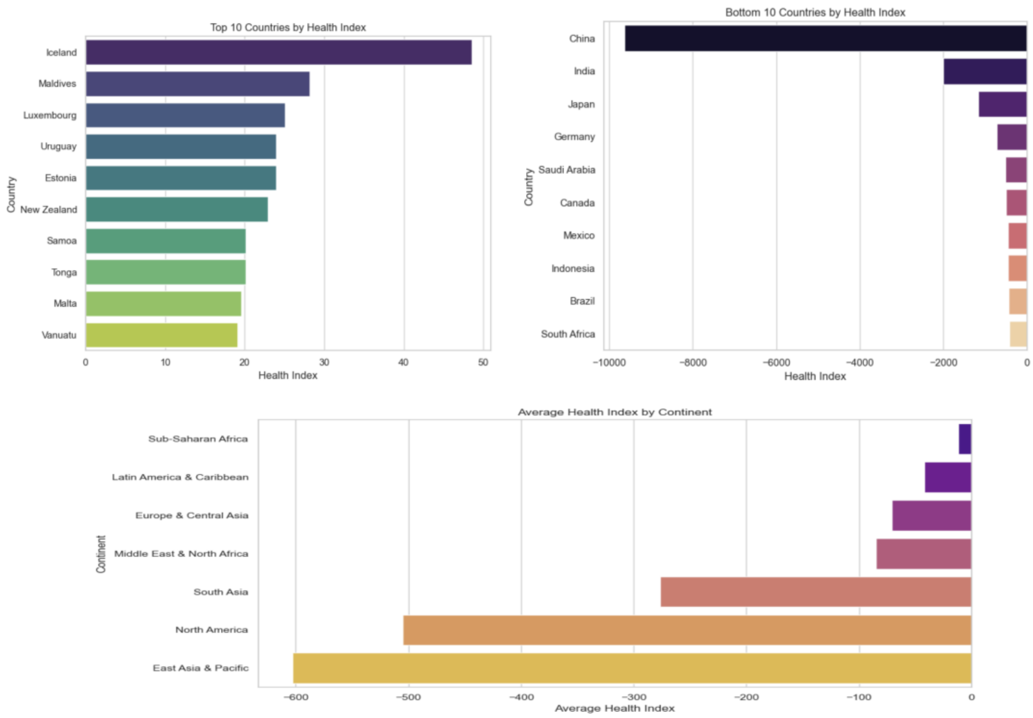
The line chart on the left illustrates how **average CO_2 emissions evolved over time across continents** up to 2015. North America consistently exhibits the highest emission levels, although there is a slight downward trend from 2010 to 2015. In contrast, East Asia & Pacific and South Asia show gradual increases, reflecting the impact of industrialization and economic growth in those regions. Other continents remain relatively stable with comparatively lower emission levels.

The scatter plot on the right explores the relationship between economic development (GDP) and CO_2 emissions across countries. Using logarithmic scales, it reveals a positive correlation, which countries with higher GDP generally tend to produce more CO_2 emissions. Some high-GDP countries show relatively lower emissions, suggesting that wealthier economies may adopt cleaner technologies or stricter environmental policies. Conversely, many lower-GDP countries still exhibit high emissions relative to their economic output, underscoring issues of carbon intensity in developing regions.

Taken together, these two graphs demonstrate that while CO_2 emissions are strongly linked to economic growth, the relationship varies significantly by region and level of development.

Research Question 5

How do countries and continents perform on the Composite Health–Economy–Environment Index based on formula we designed?



Interpretation

To provide an integrated measure that captures the trade-offs between public health, environmental quality, and economic development, we construct a **Composite Health–Economy–Environment Index (CHEE Index)**. The formula is:

$$CHEE\ Index = \frac{Life\ expectancy}{PM_{2.5}} \times ln(GDPpercapita + 1) - \frac{CO_2\ missions}{1000}$$

- **Life Expectancy / PM2.5**: This ratio reflects the balance between health outcomes and air quality. Higher life expectancy combined with lower pollution raises the index.

- **ln(GDP per capita + 1)**: Economic development is included, but the logarithmic transformation reduces the impact of extremely high-income countries and emphasizes proportional gains.
- **CO_2 Emissions / 1000**: A penalty term is applied for high carbon emissions, ensuring that unsustainable growth lowers the index.

The first chart (Top 10 Countries by Health Index) highlights the nations that achieve the best balance between health, economy, and environment. Countries such as **Iceland, Maldives, and Luxembourg** stand out with the highest scores, suggesting strong public health outcomes, manageable environmental conditions, and relatively sustainable economic performance. Other small or high-income countries (e.g., **Uruguay, New Zealand, Malta**) also perform well, which may indicate that both scale and policy choices contribute to higher resilience.

In contrast, the second chart (Bottom 10 Countries by Health Index) reveals countries with the lowest scores. Surprisingly, some large economies such as **China, India, Japan, and Germany** appear at the bottom, reflecting how high CO_2 emissions and pollution burdens can offset gains in life expectancy and GDP. This underscores that economic strength alone does not guarantee better outcomes when environmental pressures are factored in.

The third chart (Average Health Index by Continent) compares continents. **Sub-Saharan Africa and Latin America & Caribbean** show relatively stronger average values, while **East Asia & Pacific and North America** fall to the lowest levels due to the environmental costs of rapid industrialization and high consumption. This continental perspective reinforces the disparities revealed at the country level, illustrating how regional development models influence the balance between health, economy, and the environment.

Insight and Future Work

This project highlights the complex interplay between economic development, environmental quality, and public health across countries. The analysis shows that higher pollution levels, particularly PM2.5, are generally associated with lower life expectancy, though the impact varies depending on regional policies and healthcare capacity. Economic growth does not necessarily come at the expense of the environment—some nations achieve both high GDP and good air quality through effective regulations and cleaner technologies. CO_2 emissions remain closely linked to development but can be mitigated with sustainable practices. The CHEE Index further identifies countries that effectively balance health, environmental quality, and economic performance.

Future work could include:

1. **Expanded Variables:** Adding factors such as healthcare access, education, or urbanization to better understand drivers of health and environmental outcomes.
2. **Temporal and Spatial Analysis:** Studying trends over time or at regional levels to identify local challenges and policy successes.
3. **Predictive Modelling:** Developing models to forecast life expectancy or environmental impacts under different economic and policy scenarios.
4. **Policy Evaluation:** Assessing the effectiveness of specific interventions to inform sustainable development strategies.

References

ChatGPT. (2025). OpenAI. https://chat.openai.com

Gene M. Grossman, & Alan B. Krueger. (1995). *Economic Growth and the Environment.* https://doi.org/10.2307/2118443

Qiang Wang & Min Su. (2020). *Drivers of decoupling economic growth from carbon emission – an empirical analysis of 192 countries using decoupling model and decomposition method*. https://doi.org/10.1016/j.eiar.2019.106356

Richard Burnett, Hong Chen, Mieczysław Szyszkowicz, & Joseph V. Spadaro. (2018). *Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter*. https://doi.org/10.1073/pnas.1803222115

Appendix A
Data Source

Dataset 1

Source name: Life Expectancy Data
Source type: CSV
Source size: 21columns and 2865 rows
Source description: This dataset contains country-level life expectancy data from 2000 to 2015, compiled by the World Health Organization (WHO). It includes both numeric and categorical variables, such as life expectancy (in years), adult mortality rate, infant deaths, health expenditure, schooling, GDP per capita, and other demographic and health-related indicators. Each row represents a country-year observation, allowing for time-series and cross-country analysis of public health trends.
Source link: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Dataset 2

Source name: PM2.5 Global Air Pollution 2010-2017
Source type: CSV
Source size: 10 columns and 241 rows
Source description: This dataset provides annual average PM2.5 air pollution levels for countries worldwide from 2010 to 2017. It includes both numeric and categorical variables, such as country name, country code, continent, region, year, and the average PM2.5 concentration (in micrograms per cubic meter). The data allows for temporal and geographical analysis of air quality trends and can be linked with socio-economic and health datasets for further study.
Source link: https://www.kaggle.com/datasets/kweinmeister/pm25-global-airpollution-20102017

Dataset 3

Source name: World GDP by Country, Region, and Income Group

Source type: CSV

Source size: 3 columns and 218 rows

Source description: This dataset contains information on countries worldwide, including their ISO country code, geographical region, and income group classification. The data can be used for comparative analysis across regions, income levels, and for linking with other socio-economic or environmental datasets.

Source link: https://www.kaggle.com/datasets/sazidthe1/world-gdp-data

Dataset 4

Source name: World GDP by Country, Region, and Income Group

Source type: CSV

Source size: 3 columns and 217 rows

Source description: This dataset comprises various key indicators related to GDP, covering from 1960 up to 2022.

Source link: https://www.kaggle.com/datasets/sazidthe1/world-gdp-data

Dataset 5

Source name: Life expectancy & Socio-Economic (world bank)

Source type: CSV

Source size: 16 columns and 3306 rows

Source description: This dataset has shown how life expectancy varies between high-income and low-income countries, and the relationship with the CO2 emissions.

Source link: https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank?resource=download

Appendix B
SQL Statements

Question 1 Query

```
SELECT
   l.continent_name,
   l.country_name,
   round(AVG(CASE WHEN i.indicator_name='PM2.5' THEN round(f.value,2)
END),2) AS avg_pm25,
   round(AVG(CASE WHEN i.indicator_name='Life_expectancy' THEN
round(f.value,2) END),2) AS avg_life
FROM Fact_Metrics f
JOIN Dim_Location l ON f.location_id = l.location_id
JOIN Dim_Indicator i ON f.indicator_id = i.indicator_id
GROUP BY l.continent_name, l.country_name
ORDER BY l.continent_name, avg_life DESC;
```

continent_name	country_name	avg_pm25	avg_life
East Asia & Pacific	New Zealand	6.81	83.42
East Asia & Pacific	Japan	13.34	83.25
East Asia & Pacific	Singapore	19.61	82.57
East Asia & Pacific	Australia	10.17	82.37
East Asia & Pacific	China	64.7	75.52
East Asia & Pacific	Malaysia	18.15	74.55
East Asia & Pacific	Thailand	30.43	74.38
East Asia & Pacific	Samoa	13.49	73.37

Q1 Visualisation Code (Python)

```
df = pd.read_csv("Q1.csv")



df_continent = df.groupby("Continent", as_index=False).agg({
    "avg_pm25": "mean",
    "avg_life_expectancy": "mean"
})


plt.figure(figsize=(9,6))
sns.scatterplot(
    data=df_continent,
    x="avg_pm25",
```

```
    y="avg_life_expectency",
    hue="Continent",
    s=200,
    palette="Set2",
    edgecolor="black"
)
for i, row in df_continent.iterrows():
    plt.text(row["avg_pm25"]+0.5, row["avg_life_expectency"]+0.5, row["Continent"], fontsize=9)
plt.title("PM2.5 vs Life Expectancy (Continent-level)")
plt.xlabel("Average PM2.5 (µg/m³)")
plt.ylabel("Average Life Expectancy (years)")
plt.show()


df_extremes = df.loc[df.groupby("Continent")["avg_pm25"].idxmax()]
df_extremes["type"] = "Max PM2.5"
df_min = df.loc[df.groupby("Continent")["avg_pm25"].idxmin()]
df_min["type"] = "Min PM2.5"

df_extremes = pd.concat([df_extremes, df_min])

plt.figure(figsize=(9,6))
sns.barplot(
    data=df_extremes,
    x="Continent",
    y="avg_pm25",
    hue="type",
    palette="coolwarm"
)
plt.title("Countries with Max/Min PM2.5 within Each Continent")
plt.ylabel("Average PM2.5 (µg/m³)")
plt.xticks(rotation=30)
plt.show()
```

Question 2 Query

```
SELECT
    l.continent_name,
    l.country_name,
    y.year,
```

```sql
    AVG(CASE WHEN i.indicator_name='GDP_per_capita' THEN f.value END) AS
avg_gdp,
    AVG(CASE WHEN i.indicator_name='PM2.5' THEN f.value END) AS avg_pm25
FROM Fact_Metrics f
JOIN Dim_Location l ON f.location_id = l.location_id
JOIN Dim_Indicator i ON f.indicator_id = i.indicator_id
JOIN Dim_Year y ON f.year_id = y.year_id
GROUP BY l.continent_name, l.country_name, y.year
ORDER BY l.continent_name, l.country_name, y.year;
```

continent_name	country_name	year	avg_gdp	avg_pm25
East Asia & Pacific	American Samoa	2010	NULL	15.07
East Asia & Pacific	American Samoa	2011	NULL	15.39
East Asia & Pacific	American Samoa	2012	NULL	14.34
East Asia & Pacific	American Samoa	2013	NULL	14.13
East Asia & Pacific	American Samoa	2014	NULL	13.32
East Asia & Pacific	American Samoa	2015	NULL	13.02
East Asia & Pacific	Australia	2010	51874.85	10.62
East Asia & Pacific	Australia	2011	62245.13	11.05

Q2 Visualisation Code (Python)

```python
df = pd.read_csv("Q2.csv")


df = df.dropna(subset=["avg_gdp"])


plt.figure(figsize=(8,6))
sns.lineplot(data=df, x="year", y="avg_gdp", hue="Continent", marker="o")
plt.title("GDP per Capita Trend over Time by Continent")
plt.ylabel("Average GDP per Capita")
plt.show()


plt.figure(figsize=(8,6))
sns.lineplot(data=df, x="year", y="avg_pm25", hue="Continent", marker="o")
plt.title("PM2.5 Trend over Time by Continent")
plt.ylabel("Average PM2.5 (µg/m³)")
```

```python
plt.show()



plt.figure(figsize=(8,6))
sns.scatterplot(
    data=df,
    x="avg_gdp",
    y="avg_pm25",
    hue="Continent",
    size="year",
    sizes=(40,200),
    alpha=0.7
)
plt.title("GDP vs PM2.5 (Bubble Size = Year)")
plt.xlabel("Average GDP per Capita")
plt.ylabel("Average PM2.5 (µg/m³)")
plt.legend(bbox_to_anchor=(1.05, 1), loc=2)
plt.show()
```

Question 3 Query

```sql
SELECT
    l.continent_name,
    l.country_name,
    b.band_name,
    COUNT(*) AS n_records
FROM Fact_Metrics f
JOIN Dim_Location l ON f.location_id = l.location_id
JOIN Dim_PM25Band b ON f.band_id = b.band_id
GROUP BY l.continent_name, l.country_name, b.band_name
ORDER BY b.band_name, l.continent_name;
```

continent_name	country_name	band_name	n_records
East Asia & Pacific	Australia	Good	3
East Asia & Pacific	New Zealand	Good	6
Europe & Central Asia	Estonia	Good	6
Europe & Central Asia	Finland	Good	6
Europe & Central Asia	Iceland	Good	6
Europe & Central Asia	Ireland	Good	5
Europe & Central Asia	Norway	Good	6
Europe & Central Asia	Portugal	Good	4

Q3 Visualisation Code (Python)

```
file_path = "Q3.csv"
data = pd.read_csv(file_path)

sns.set(style="whitegrid")


continent_summary = data.groupby(["Continent", "band_level"])["Record"].sum().unstack(fill_value=0)

ax1 = continent_summary.plot(kind="bar", stacked=True, figsize=(8,6))
plt.title("Distribution of Band Levels by Continent")
plt.ylabel("Total Records")
plt.xlabel("Continent")
plt.xticks(rotation=45, ha="right")
plt.legend(title="Band Level")
plt.tight_layout()
plt.show()


band_summary = data.groupby("band_level")["Record"].sum()

plt.figure(figsize=(6,6))
plt.pie(band_summary, labels=band_summary.index, autopct="%1.1f%%", startangle=140)
plt.title("Overall Band Level Distribution")
plt.tight_layout()
plt.show()
```

Question 4 Query

```
WITH CO2 AS (
    SELECT f.location_id, f.year_id, AVG(f.value) AS avg_CO2
    FROM Fact_Metrics f
    JOIN Dim_Indicator i ON f.indicator_id = i.indicator_id
    WHERE i.indicator_name = 'CO2_emissions'
    GROUP BY f.location_id, f.year_id
),
gdp AS (
    SELECT f.location_id, f.year_id, AVG(f.value) AS avg_gdp
    FROM Fact_Metrics f
    JOIN Dim_Indicator i ON f.indicator_id = i.indicator_id
    WHERE i.indicator_name = 'GDP_per_capita'
    GROUP BY f.location_id, f.year_id
)
SELECT
    l.continent_name,
    l.country_name,
    y.year,
    CO2.avg_CO2,
    gdp.avg_gdp,
    round(CO2.avg_CO2 / NULLIF(gdp.avg_gdp,0),2) AS CO2_per_gdp
FROM Dim_Location l
JOIN Dim_Year y ON 1=1
LEFT JOIN CO2 ON l.location_id=CO2.location_id AND y.year_id=CO2.year_id
LEFT JOIN gdp ON l.location_id=gdp.location_id AND y.year_id=gdp.year_id
ORDER BY l.continent_name, l.country_name, y.year;
```

continent_n...	country_name	year	avg_co2	avg_gdp	co2_per_gdp
East Asia & Pacific	American Samoa	2010	0	NULL	NULL
East Asia & Pacific	American Samoa	2011	0	NULL	NULL
East Asia & Pacific	American Samoa	2012	0	NULL	NULL
East Asia & Pacific	American Samoa	2013	0	NULL	NULL
East Asia & Pacific	American Samoa	2014	0	NULL	NULL
East Asia & Pacific	American Samoa	2015	0	NULL	NULL
East Asia & Pacific	Australia	2010	387540.01	51874.85	7.47
East Asia & Pacific	Australia	2011	386380	62245.13	6.21

Q4 Visualisation Code (Python)

```python
q4_data = pd.read_csv("Q4.csv")
```

```
sns.set(style="whitegrid")


q4_data = q4_data[q4_data["year"] <= 2015]


plt.figure(figsize=(8,6))
trend = q4_data.groupby(["year","Continent"])["avg_co2"].mean().reset_index()


sns.lineplot(data=trend, x="year", y="avg_co2", hue="Continent", marker="o")
plt.title("Average CO2 Emissions by Continent (≤2015)")
plt.ylabel("Average CO2 Emissions")
plt.xlabel("Year")
plt.legend(title="Continent")
plt.tight_layout()
plt.show()



plt.figure(figsize=(8,6))


scatter_data = q4_data.dropna(subset=["avg_co2", "avg_gdp"])
sns.scatterplot(data=scatter_data, x="avg_gdp", y="avg_co2", hue="Continent", alpha=0.7)


plt.xscale("log")
plt.yscale("log")
plt.title("Relationship Between GDP and CO2 Emissions (≤2015)")
plt.xlabel("Average GDP (log scale)")
plt.ylabel("Average CO2 Emissions (log scale)")
plt.legend(title="Continent")
plt.tight_layout()
plt.show()
```

Question 5 Query

```sql
WITH country_avgs AS (
   SELECT
      l.continent_name,
      l.country_name,
      AVG(CASE WHEN i.indicator_name='Life_expectancy' THEN f.value END) AS
avg_life,
      AVG(CASE WHEN i.indicator_name='PM2.5' THEN f.value END) AS avg_pm25,
      AVG(CASE WHEN i.indicator_name='GDP_per_capita' THEN f.value END) AS
avg_gdp,
      AVG(CASE WHEN i.indicator_name='CO2_emissions' THEN f.value END) AS
avg_CO2
   FROM Fact_Metrics f
   JOIN Dim_Location l ON f.location_id = l.location_id
   JOIN Dim_Indicator i ON f.indicator_id = i.indicator_id
   GROUP BY l.continent_name, l.country_name
)
SELECT
   continent_name,
   country_name,
   round((avg_life / avg_pm25) * log(avg_gdp + 1) - (avg_CO2/1000),2) AS health_index
FROM country_avgs
ORDER BY health_index DESC;
```

continent_name	country_name	health_index
Europe & Central Asia	Iceland	46.98
South Asia	Maldives	26.38
Europe & Central Asia	Luxembourg	23.63
Latin America & Caribbean	Uruguay	22.81
Europe & Central Asia	Estonia	22.2
East Asia & Pacific	New Zealand	21.37
East Asia & Pacific	Tonga	19.47
East Asia & Pacific	Samoa	19.4

Q5 Visualisation Code (Python)

```python
q5_data = pd.read_csv("Q5.csv")


q5_data = q5_data.dropna(subset=["Health_index"])


top_countries = q5_data.sort_values("Health_index", ascending=False).head(10)


plt.figure(figsize=(8,6))
sns.barplot(data=top_countries, x="Health_index", y="Country", palette="viridis")
plt.title("Top 10 Countries by Health Index")
```

```
plt.xlabel("Health Index")
plt.ylabel("Country")
plt.tight_layout()
plt.show()

bottom_countries = q5_data.sort_values("Health_index", ascending=True).head(10)

plt.figure(figsize=(8,6))
sns.barplot(data=bottom_countries, x="Health_index", y="Country", palette="magma")
plt.title("Bottom 10 Countries by Health Index")
plt.xlabel("Health Index")
plt.ylabel("Country")
plt.tight_layout()
plt.show()

continent_avg = q5_data.groupby("Continent")["Health_index"].mean().reset_index()

plt.figure(figsize=(12,6))
sns.barplot(data=continent_avg.sort_values("Health_index", ascending=False),
            x="Health_index", y="Continent", palette="plasma")
plt.title("Average Health Index by Continent")
plt.xlabel("Average Health Index")
plt.ylabel("Continent")
plt.tight_layout()
plt.show()
```