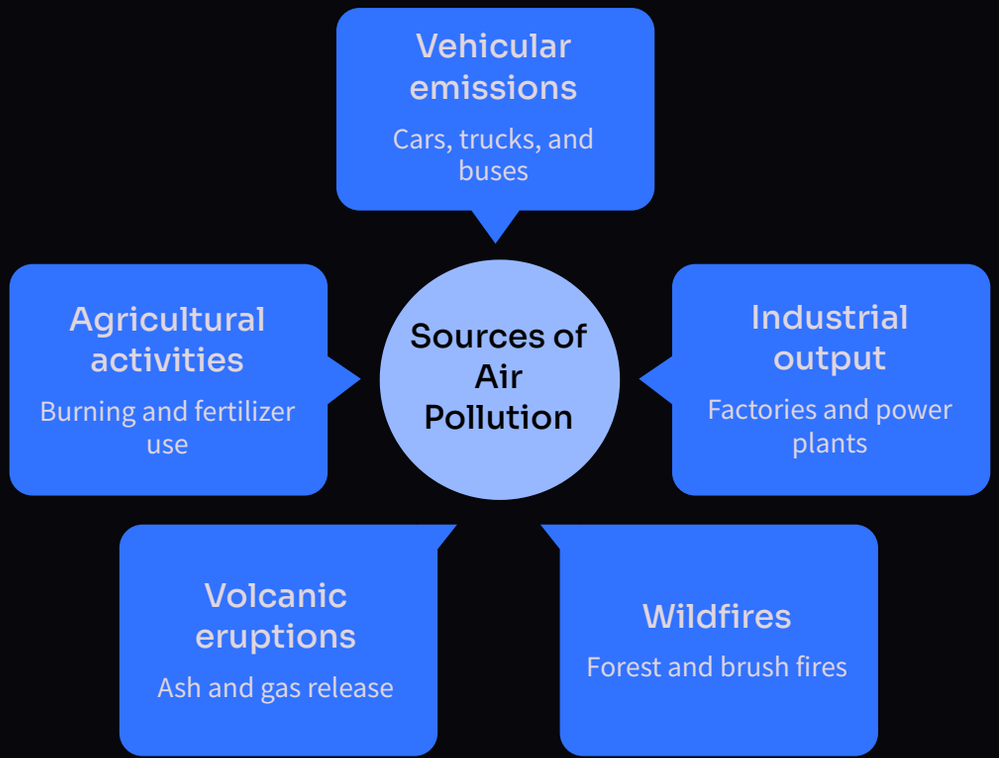
Air Pollution: A Global Crisis Demanding Action

by Rawan Ahmed

Full project on GitHub

What Is Air Pollution?

Air pollution is the introduction of harmful substances such as tiny particulates, toxic gases, and biological molecules into Earth's atmosphere.



Made with GAMMA

The Deadly Scale: 8.1 Million Premature Deaths in 2021

8.1M

Lives Lost

Air pollution caused more deaths than HIV/AIDS and malaria combined in 2021.

700K+

Children Affected

Over 700,000 children under five died due to exposure to polluted air.

99%

Global Exposure

Virtually the entire global population breathes air exceeding WHO pollution limits.

Made with GAMMA

Key Pollutants to Know

Particulate Matter (PM2.5 & PM10)

Tiny particles that can penetrate deep into the lungs and bloodstream, causing severe health issues.

Nitrogen Oxides (NOx) & Sulfur Dioxide (SO2)

Primarily from fossil fuel combustion, these lead to smog formation and acid rain, harming ecosystems.

Carbon Monoxide (CO)

An odorless, toxic gas produced by incomplete combustion, highly dangerous in enclosed spaces.

Volatile Organic Compounds (VOCs)

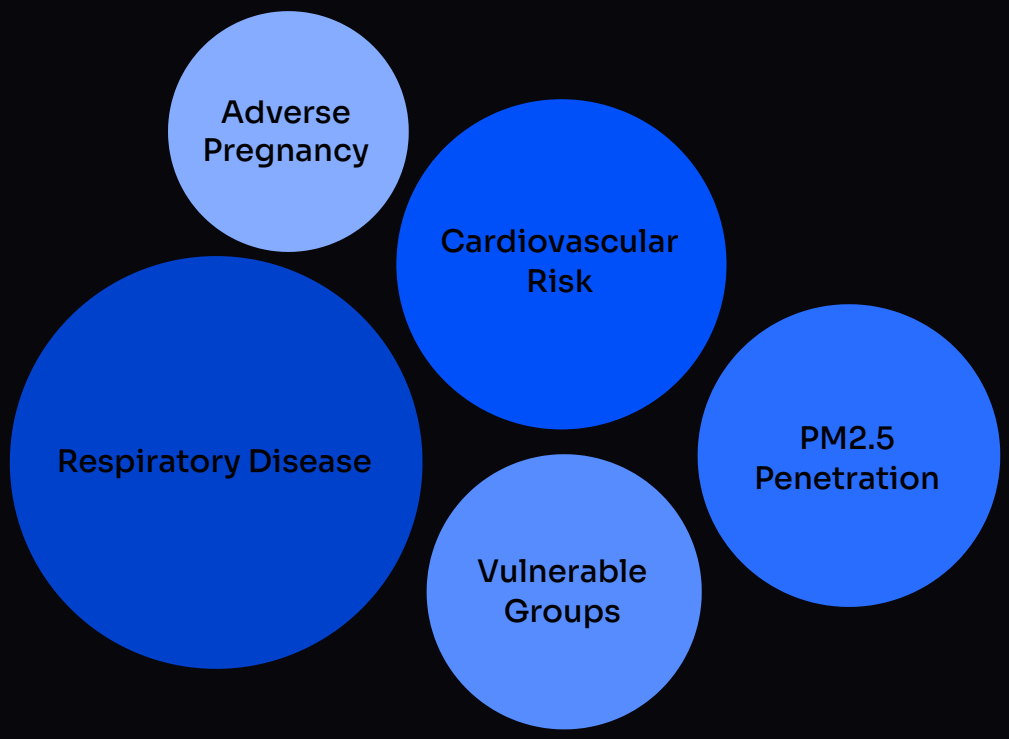
Carcinogenic chemicals emitted from industrial processes and common household products.

Made with GAMMA

Health Impacts: More Than Just a Cough

Air pollution's effects extend far beyond respiratory issues. It is strongly linked to serious conditions such as pneumonia, asthma, lung cancer, and increases the risk of heart disease and stroke.

Fine particulate matter **(PM2.5)** is particularly dangerous as it can enter the bloodstream, affecting multiple organs and body systems. Children, the elderly, and individuals with pre-existing conditions are the most vulnerable to long-term damage, facing chronic illnesses and even premature death. It also contributes to adverse pregnancy outcomes like stillbirths.



Made with GAMMA

Dataset overview:

Data Gathering By: Rawan Ahmed

AirQulaity Dataset summary:

Source of Data: WHO (world health organization).

- This data set provides a snapshot of air quality metrics and demographic information.
- Geographical Focus.
- Time Reference.
- Air Pollutant Concentration.

Health Dataset summary:

Source of Data: IHME (Global Burden of disease).

- This data set appears to contain global or multi-country health statistics.
- Primary Focus (Measure).
- Geographical Scope.
- Observed Values.

Made with GAMMA

Data Cleaning:

Data Cleaning by: Rawan Ahmed and Ibrahim Abdelmohsen

AirQuality Data :

```
AirQuality.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
Index: 40095 entries, 0 to 40097
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   who_region          40095 non-null  object
 1   iso3                40095 non-null  object
 2   country_name        40095 non-null  object
 3   city                40095 non-null  object
 4   year                40095 non-null  int64
 5   version             40095 non-null  object
 6   pm10_concentration  40095 non-null  float64
 7   pm25_concentration  40095 non-null  float64
 8   no2_concentration   40095 non-null  float64
 9   pm10_tempcov        40095 non-null  float64
 10  pm25_tempcov        40095 non-null  float64
 11  no2_tempcov         40095 non-null  float64
 12  type_of_stations    40095 non-null  object
 13  number_of_stations  40095 non-null  int64
 14  population          40095 non-null  float64
 15  population_source   40095 non-null  object
 16  latitude            40095 non-null  float64
 17  longitude           40095 non-null  float64
dtypes: float64(9), int64(2), object(7)
memory usage: 5.8+ MB
```

```
AirQuality.isnull().sum()
```

```
who_region            0
iso3                  0
country_name          0
city                  0
year                  0
version               0
pm10_concentration    0
pm25_concentration    0
no2_concentration     0
pm10_tempcov          0
pm25_tempcov          0
no2_tempcov           0
type_of_stations      0
number_of_stations    0
population            0
population_source     0
latitude              0
longitude             0
dtype: int64
```

Made with GAMMA

Data Cleaning:

Health Data

```
#detect duplicates:
health_data.duplicated()
```

	0
0	False
1	False
2	False
3	False
4	False
...	...
9823	False
9824	False
9825	False
9826	False
9827	False

9828 rows × 1 columns

dtype: bool

all the id columns are unncessary because their names are existing in the data

```
health_data= health_data.drop(columns=[
    'measure_id',
    'location_id',
    'sex_id',
    'age_id',
    'cause_id',
    'metric_id'
])
```

Change Names of some Columns to match other data sets:

```
health_data = health_data.rename(columns={
    'location_name': 'country',
    'measure_name': 'measure',
    'cause_name': 'disease',
    'val': 'value'
})
```

Made with GAMMA

SQL Analysis (Using SQL Server)

Analysis by: Ghalia Hamed and Mohamed Sayed

"After cleaning and preparing the dataset, we moved to the SQL analysis phase using SQL Server to answer key questions about air pollution trends and health impact."

▪ Top 5 Worst Air-Quality Countries (Annually)

We identified the top 5 countries with the highest PM2.5 concentrations each year.
This helps highlight the most polluted regions annually and shows how rankings change over time.

```sql
--4)Top 5 worst air-quality countries annually
WITH aggregated AS (
    SELECT
        country_name,
        year,
        AVG(pm25_concentration) AS avg_pm25
    FROM AirQuality_cleaned2
    GROUP BY country_name, year
),
ranked AS (
    SELECT
        country_name,
        year,
        avg_pm25,
        DENSE_RANK() OVER (PARTITION BY year ORDER BY avg_pm25 DESC) AS pollution_rank
    FROM aggregated
)
SELECT *
FROM ranked
WHERE pollution_rank <= 5;
```

Made with GAMMA

SQL Analysis (Using SQL Server)

▪ Did deaths rise or fall?

This SQL query calculates the **Year-over-Year (YoY) change in total deaths** for each country using a Common Table Expression (CTE) to aggregate the data first. The **LAG window function** is then applied to retrieve the total death figure from the preceding year for comparison.

```sql
--6)Did deaths rise or fall?
WITH deaths_agg AS (
    SELECT
        country,
        year,
        SUM(value) AS total_deaths
    FROM [cleaned_health_data-1]
    GROUP BY country, year
)
SELECT
    country,
    year,
    total_deaths,
    LAG(total_deaths) OVER (PARTITION BY country ORDER BY year) AS prev_year_deaths,
    total_deaths - LAG(total_deaths) OVER (PARTITION BY country ORDER BY year) AS yoy_change_deaths
FROM deaths_agg
ORDER BY country, year;
```

Made with GAMMA

SQL Analysis (Using SQL Server)

▪Top 3 Pollution-Related Mortality Leaders (Per Region)

We identified the three countries with the highest pollution-related death rates in each WHO region. This helps highlight the main hotspots within every region and shows where the health impact of air pollution is most severe.

```sql
--12)Which age group has the highest average death count across all countries and years.
SELECT
    age,
    AVG(value) AS avg_deaths,
    SUM(value) AS total_deaths
FROM [cleaned_health_data-1]
GROUP BY age
ORDER BY avg_deaths DESC;
```

Made with GAMMA

SQL Analysis (Using SQL Server)

▪Which countries have the highest long-term pollution risk score?

This SQL query identifies countries with the **highest long-term pollution risk** by calculating a yearly risk score (PM2.5 concentration times population). It then averages these yearly scores to determine the overall **long-term risk** for each country and ranks them in descending order.

```sql
--15)Which countries have the highest long-term pollution risk score (PM2.5 * population).
WITH yearly AS (
    SELECT
        country_name,
        year,
        AVG(pm25_concentration) AS avg_pm25,
        AVG(population) AS population
    FROM AirQuality_cleaned2
    GROUP BY country_name, year
),
risk AS (
    SELECT
        country_name,
        year,
        (avg_pm25 * population) AS risk_score
    FROM yearly
)
SELECT
    country_name,
    AVG(risk_score) AS long_term_risk
FROM risk
GROUP BY country_name
ORDER BY long_term_risk DESC;
```

Made with GAMMA

DashBoard & Visualization Insights (Using Power Bi)

Created By: Abdallah Ahmed

Dashboard Overview

After completing the data cleaning and SQL analysis, we built an interactive dashboard to visualize the key insights from the air-quality dataset.
The dashboard helps us clearly understand pollution patterns, category distribution, and health impacts across countries.

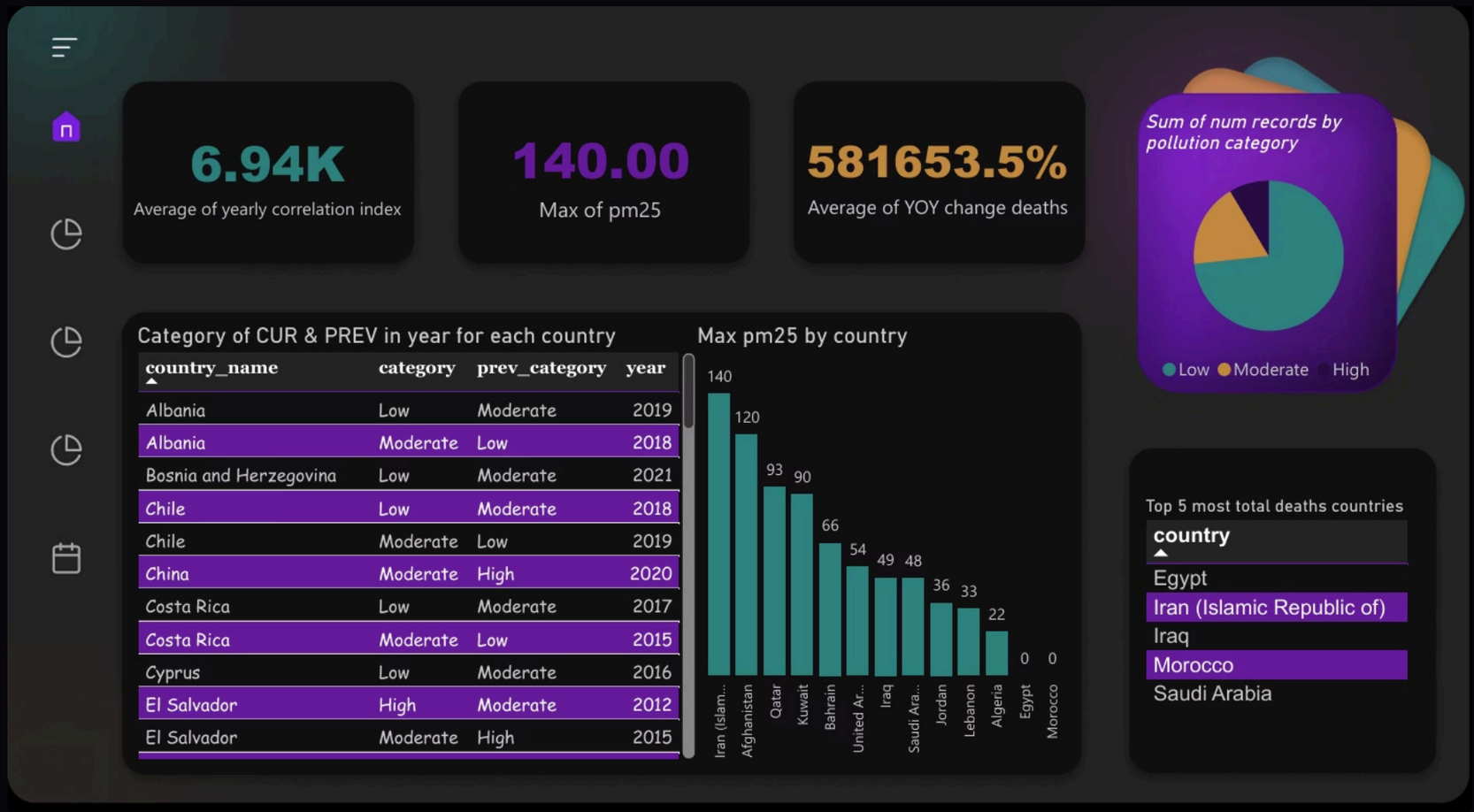
Dashboard Purpose

The dashboard transforms raw numbers into easy-to-understand visuals, allowing us to:

- Identify high-risk countries
- Monitor pollution trends
- Detect category shifts over time
- Link pollution with health impact

Made with GAMMA

DashBoard Details



Key Metrics Cards

Pollution Category Distribution

Category Change per Country

Max PM2.5 by Country

Top 5 Countries with Highest Total Deaths

Made with GAMMA

DashBoard Details

Key Health and Environmental Metrics

Country Contribution to Improvement

Mortality by Age Group

PM2.5 or Risk Index Trend

Focus on Environmental Risk

Made with GAMMA

DashBoard Details

Made with GAMMA

Key Findings

Air pollution levels vary significantly across countries and years

PM2.5 is the most critical pollutant and reaches extremely high values in certain regions

Several countries showed year-to-year improvement, while others experienced continuous decline.

The 3-year rolling trend helped reveal long-term patterns beyond annual fluctuations.

Pollution-related deaths remain highest in a small group of countries, emphasizing the health risk.

Made with GAMMA

Tools Used:

Python

Versatile scripting and data processing

Jupyter

Interactive notebooks for analysis

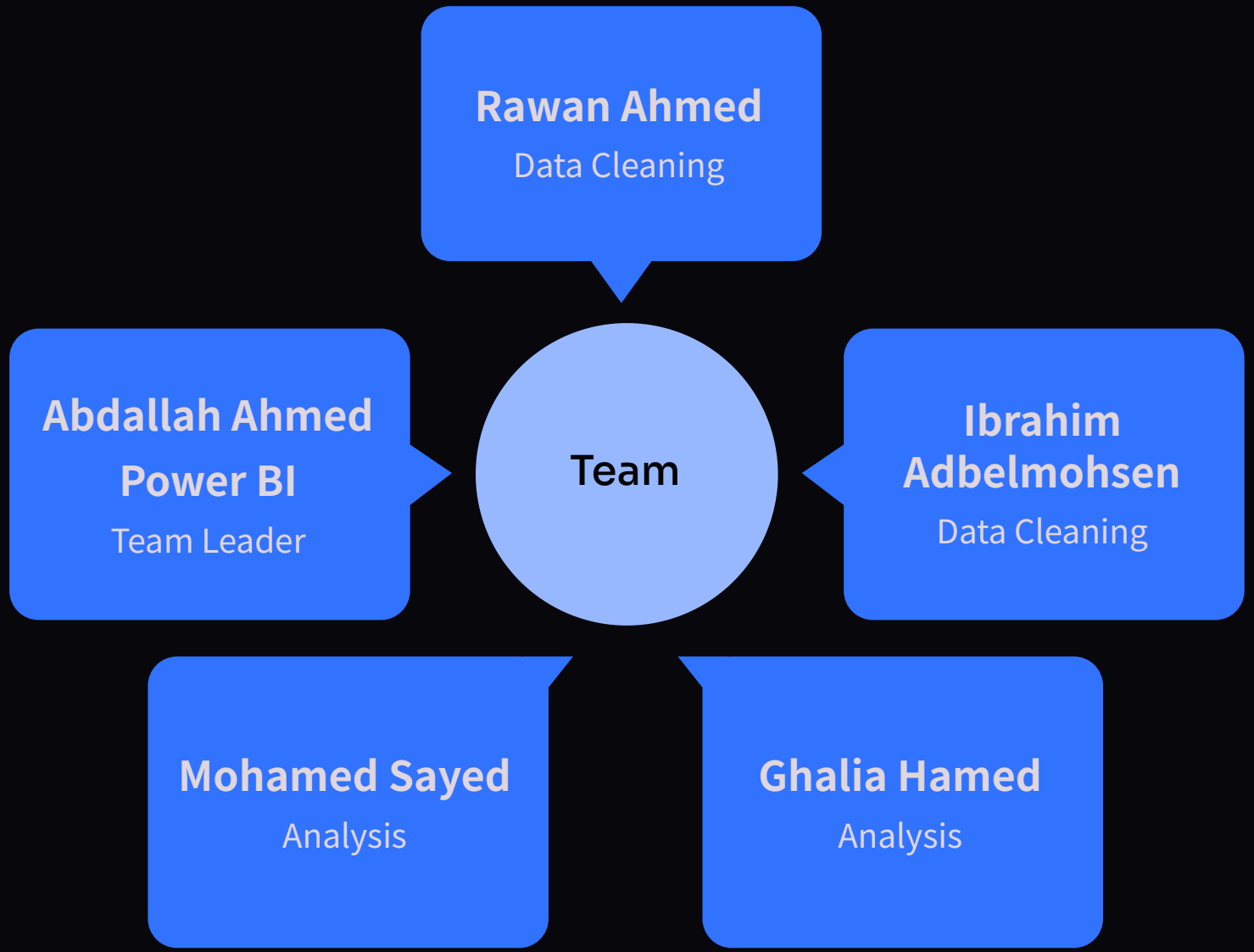
SQL Server

Reliable relational database engine

Power BI

Visualize insights and dashboards

Made with GAMMA



Made with GAMMA