AIR POLLUTION

PROJECT DOCUMENTATION

BY: Rawan Ahmed

TABLE OF CONTENTS:

INTRODUCTION

Air pollution is one of the most critical environmental problems in the modern world. It directly affects human health, contributes to climate change, and increases the risk of respiratory and cardiovascular diseases.

This project aims to analyze global air pollution levels, specifically focusing on PM2.5 concentration and its relationship to death rates across different countries and WHO regions.

The entire workflow includes three major stages:

- Data cleaning using Python
- Data storage and analysis using SQL Server
- Developing an interactive Power BI dashboard for visualization and reporting

This documentation describes each step in detail, explaining how the raw data was transformed into meaningful insights.

1.BACKGROUND & PROBLEM DEFINATION

The quality of air varies greatly between countries due to differences in industrialization, population density, climate conditions, and environmental policies.

Poor air quality can lead to serious health consequences, especially in low- and middle-income countries where pollution control systems are limited.

The dataset used in this project contains real-world air quality and health metrics. However, raw data often includes errors, missing values, or inconsistencies. If not cleaned properly, it can lead to inaccurate analyses and misleading conclusions.

Therefore, the main problem this project solves is:

"How can we clean, analyze, and visualize global air pollution data to identify the most affected countries and regions, and understand the relationship between pollution and death rates?"

2. PROJECT OBJECTIVES

The main objectives of the project are:

- To perform thorough data cleaning to ensure accuracy and reliability.
- To identify countries with the highest PM2.5 levels annually.
- To calculate a simple correlation between PM2.5 values and death rates.
- To determine the top 3 countries with the highest death rates within each WHO region.
- To create an interactive dashboard in Power BI that summarizes all insights.
- To provide a clear and structured report that explains both methodology and results.

3.DATA SET DESCRIPTION

The dataset includes global indicators related to air pollution and population health.

It contains detailed records for multiple countries over different years.

Key Columns:

- Country Name: Name of each country in the dataset
- WHO Region: Classification of countries by World Health Organization
- PM2.5 Levels: Concentration of fine particulate matter in the air
- Death Rate: Number of deaths attributed to pollution
- Population: Total population of each country
- Year: Year of observation

Initial Data Issues Found:

- Missing values in PM2.5 and death rate columns
- Inconsistent country spellings
- Mixed data types (numeric data stored as text)
- Duplicated entries

4. TOOLS & TECHNOLOGIES

Python

Used for reading, cleaning, transforming, and exporting the dataset.

Libraries used:

- Pandas

SQL Server

Used to store the cleaned data and perform:

- Grouping
- Aggregations
- Ranking
- Filtering
- Region-based analysis

Power BI

Used to design a visually attractive and highly interactive dashboard.

Visuals included bar charts, KPIs, cards, and regional breakdowns.

5.DATA CLEANING PROCESS

Data cleaning was mainly carried out using Python. The process included:

5.1 Handling Missing Values

- Checked for missing values using ".isnull()".

- For columns like "death_rate", rows with too many missing values were removed.

- PM2.5 missing values were either filled with median values or removed depending on the severity.

5.2 Correcting Country Names

- Standardized names (e.g., “United States”, “USA”, “U.S.” → “United States of America”).

- Ensured consistency to avoid duplication during grouping.

5.3 Fixing Data Types

- Converted PM2.5 and death rate columns to numeric formats.

- Stripping symbols and unnecessary spaces.

- Ensured region column is categorical for easier grouping.

5.4 Removing Duplicates

- Detected duplicates using "df.duplicated()".

- Removed redundant records to avoid wrong aggregation.

5.5 Detecting Outliers

- Identified outliers using statistical methods like IQR.
- Decided whether to keep or remove them based on context.

5.6 Exporting Cleaned Data

After cleaning, the dataset was exported and loaded into SQL Server for deeper analysis.

6.EXPLORATORY DATA ANALYSIS

6.1Data was grouped by country

- Average PM2.5 was calculated.

- Top 5 with highest values were extracted.

Finding: Countries in South Asia (such as India, Bangladesh, Pakistan) consistently ranked highest.

6.2 Simple Correlation Indicator

A correlation was calculated between PM2.5 and death rate.

Result:

A positive correlation indicates that higher PM2.5 levels are associated with higher pollution-related deaths.

6.3Top 3 Death Rates per WHO Region

Using SQL:

```
SELECT TOP 3 Country, Region, DeathRate
FROM AirPollution
ORDER BY Region, DeathRate DESC;
```

This helped identify the most affected countries within each region. Insights showed that certain regions like Africa and South-East Asia had significantly higher death rates in multiple countries.

7. POWER BI DASHBOARD

The Power BI dashboard is designed to give a clear overview of global air pollution patterns.

7.1 PM2.5 Visualization

A bar chart ranking countries from highest to lowest PM2.5 levels.

This visual makes it easy to see which countries suffer from the worst air quality.

7.2 Correlation Section

A scatter chart showing the relationship between PM2.5 and death rate.

This helps users understand how pollution impacts mortality.

7.3 WHO Region Summary

A matrix or clustered bar chart showing:

- The highest 3 death rates in each WHO region
- Comparison between different countries.

7.4 KPI Cards

Include summarized values:

- Global average PM2.5

- Total number of countries
- Highest recorded PM2.5
- Region with highest average death rate

7.5Interactive Filters

Users can filter by:

- Year
- Region
- Individual country

This makes the dashboard user-friendly and insightful.

8. KEY FINDINGS & INSIGHTS

- South Asia shows the highest levels of PM2.5 globally.
- A clear positive relationship exists between air pollution and death rates.
- Some regions, especially Africa and South-East Asia, face severe health impacts due to pollution.
- Data cleaning significantly improved the reliability of the analysis.
- Visualizations helped show major differences between countries and regions.

9. CHALLENGES FACED

- HANDLING MISSING VALUES WITHOUT LOSING IMPORTANT DATA.
- STANDARDIZING INCONSISTENT COUNTRY NAMES.
- ENSURING CORRECT JOINS IN SQL SERVER.
- CHOOSING APPROPRIATE VISUALS TO CLEARLY COMMUNICATE FINDINGS.

10. CONCLUSION

The project successfully transformed raw environmental data into a well-structured, meaningful analysis.

Through Python cleaning, SQL analysis, and Power BI visualization, the workflow demonstrated the full life cycle of a data project.

The insights provided can help raise awareness about the severity of air pollution and assist policymakers and researchers in understanding global pattern

11. FUTURE WORK

- Build predictive models to forecast future PM2.5 levels.
- Add new datasets such as GDP, health spending, or climate variables.
- Expand the dashboard with trend analysis over multiple years.
- Create machine learning models to classify countries by risk level.

TEAM:

- **ABDLLAH AHMED**
- **IBRAHIM ABDELMOHSEN**
- **MOHAMED SATED**
- **GHALIA HAMED**
- **RAWAN AHMED**