Project Phases – Air Pollution Data Analysis Project

This project follows a structured, end-to-end data lifecycle, moving from raw data to insights and dashboard reporting.
Below are the complete phases:

---

1. Project Initiation & Planning

Purpose: Define the problem, objectives, tools, and team responsibilities.
Main Activities:

- Understanding the global air pollution problem.
- Identifying the goals: PM2.5 analysis, death rates comparison, correlations, etc.
- Deciding the tools: Python, SQL Server, Power BI.
- Assigning tasks among team members.
- Identifying stakeholders.

Deliverables:

- Project outline
- Scope definition
- Stakeholder analysis (already done)

---

2. Data Collection Phase

Purpose: Obtain and understand the raw dataset.
Main Activities:

- Downloading the global pollution dataset.

- Checking dataset format (CSV/Excel).
- Understanding each column: PM2.5, death rate, year, population, WHO region.
- Identifying initial data issues.

Deliverables:

- Raw dataset
- Data description summary

---

3. Data Cleaning Phase (Python)

Purpose: Prepare accurate and reliable data for analysis.
Main Activities:

- Handling missing values (PM2.5, death rate).
- Standardizing country names.
- Converting data types (string → numeric).
- Removing duplicates.
- Detecting and handling outliers.
- Exporting cleaned data to a new file.

Deliverables:

- Cleaned dataset (CSV/Excel)
- Python scripts for data cleaning

---

4. Data Loading & Storage Phase (SQL Server)

Purpose: Store the cleaned dataset for deeper analysis.
Main Activities:

- Creating database & tables in SQL Server.
- Importing cleaned data.
- Checking for import errors.
- Ensuring correct datatypes and indexes for performance.

Deliverables:

- SQL database ready for analysis
- SQL schema documentation

---

5. Data Analysis Phase (SQL Queries)

Purpose: Extract meaningful insights using SQL analysis.
Main Activities:

- Grouping data by country and region.
- Calculating average PM2.5 levels.
- Ranking countries by death rate.
- Extracting top 3 countries per region.
- Measuring correlation (basic statistical analysis).

Deliverables:

- SQL queries
- Analysis tables/results
- Insight summary

6. Exploratory Data Analysis (EDA) Phase

Purpose: Understand trends and patterns visually.
Main Activities:

- Plotting PM2.5 distribution.
- Detecting regions with highest pollution.
- Examining correlation between PM2.5 and death rate.
- Reviewing yearly trends (if available).

Deliverables:

- EDA visuals
- Insight documentation

---

7. Dashboard Development Phase (Power BI)

Purpose: Present insights in a clear, interactive dashboard.
Main Activities:

- Importing SQL data into Power BI.
- Creating charts:
    - PM2.5 bar chart
    - Death rate matrix
    - Scatter plot for PM2.5 vs death rate
- Adding KPIs:
    - Highest PM2.5
    - Total countries
    - Region with highest death rate
- Creating slicers for year/region/country.

- Designing a clean layout for readability.

Deliverables:

- Power BI dashboard
- Dashboard documentation

---

8. Validation & Testing Phase

Purpose: Ensure accuracy and correctness of results.
Main Activities:

- Checking for visualization errors.
- Validating SQL computations.
- Reviewing Python cleaning logic.
- Getting feedback from teammates/instructor.

Deliverables:

- Validated dataset
- Corrected SQL/Python scripts
- Final refined dashboard

---

9. Reporting & Documentation Phase

Purpose: Document the full project process.
Main Activities:

- Writing project documentation (background, methodology, findings).
- Summarizing key insights.
- Preparing charts and tables.

- Writing conclusion and recommendations.

Deliverables:

- Final PDF report (your uploaded file)
- Presentation (if needed)

---

10. Presentation & Delivery Phase

Purpose: Deliver the final results.
Main Activities:

- Presenting the dashboard.
- Explaining the insights.
- Answering supervisor questions.
- Submitting final files.

Deliverables:

- Final dashboard
- Final report
- Project presentation