Friedrich-Alexander-Universität Erlangen-Nürnberg

Methods of Advanced Data Engineering

Summer Semester 2024

04.06.2024

Correlation Between World Population and Air Pollution Project Report

Berkant Cinar

Student ID: 23224469

1 Introduction

Population and air pollution are critical issues for global sustainability and public health. The goal of this study is to investigate the relationship between population size and air pollution levels across different countries. Our aim is to identify trends and correlations that provide insights into how population size impacts air quality, helping to inform policies for sustainable development and pollution control.

2 Main Question

1. Is there a direct connection between world population size and air pollution levels?

3 Data Sources

3.1 World Population Dataset

- Metadata URL: World Population Dataset
- Source: Kaggle
- Content: This dataset offers detailed information on population trends, including rates of urbanization, for all nations in the world. This dataset is really important for comprehending the size and rate of population growth in various countries.
- Data Format: CSV format with columns including 'Country', 'Continent', '2022 Population', 'Area', 'Density', 'Growth Rate', 'World Population Percentage'.
- Quality: Comprehensive dataset with structured columns but requires cleaning for missing values and standardization.

3.2 Global Air Pollution Dataset

- Metadata URL: Global Air Pollution Dataset
- Source: Kaggle
- Content: This dataset provides comprehensive data on air pollution levels, including AQI values and concentrations of various pollutants for many locations around the world. It is essential for assessing air quality and understanding the impact of different pollution sources.

- Data Format: CSV format with columns including 'Country', 'City', and various pollution metrics.
  - Nitrogen Dioxide [NO2]: Comes from vehicles, power plants, and equipment emissions. It can worsen asthma and other respiratory diseases, especially in children, the elderly, and people with asthma.
  - Ozone [O3]: Created by reactions between nitrogen oxides and volatile organic compounds in sunlight. It causes chest pain, coughing, throat irritation, and reduces lung function. It can also harm plants and ecosystems.
  - Carbon Monoxide [CO]: Emitted by vehicles and machinery burning fossil fuels, and by certain indoor appliances like gas stoves and heaters. It reduces oxygen in the bloodstream, affecting the heart and brain. High levels can cause dizziness, confusion, unconsciousness, and death, particularly in enclosed spaces.
  - Particulate Matter [PM2.5]: Tiny solid and liquid particles in the air from various sources. These particles can cause serious heart and lung problems and are classified as a carcinogen. PM2.5 particles are particularly harmful due to their small size (2.5 micrometers or less).
- Quality: Detailed dataset but includes multiple entries for the same country, requiring aggregation and cleaning.

4 Data Pipeline

4.1 Description

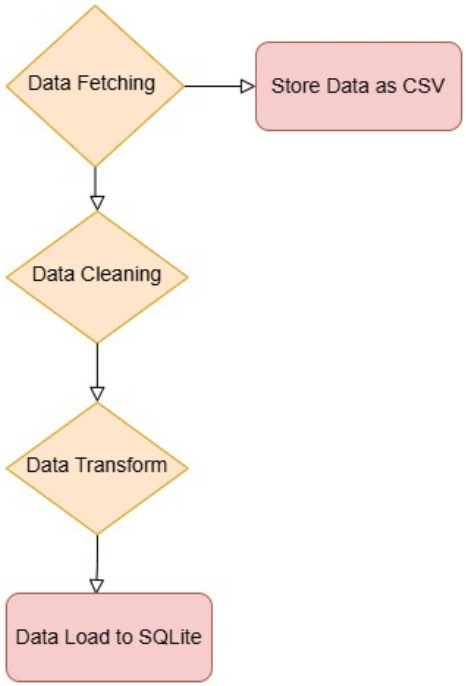
The pipeline fetches, cleans, transforms, and stores data in a SQLite database. Python, Pandas for data manipulation, SQLite for database storage, and logging for error tracking.

- Download Data: Use Kaggle API to fetch the datasets.
- Store Data as CSV: Save raw data in CSV format in the '/data' directory.
- Data Cleaning and Transform
  - **World Population Data**
    - Select relevant columns: *Country*, *Continent*, *2022 Population*, *Area*, *Density*, *Growth Rate*, *World Population Percentage*.
    - Drop rows with any null values.

    - Standardize column names.
  - **Air Pollution Data**
    - Drop the City column.
    - Drop rows with any null values.
    - Standardize column names.
    - Group by country and calculate the mean for each country.

- Data Storage: Save the processed data into an SQLite database for further analysis.

4.2 Data Pipeline Flow



5 Results and Limitations

5.1 Output Data

The pipeline output data has:

- Cleaned and transformed world population data for 2022.
- Aggregated air pollution data with mean values for each country.

First 10 Rows of World Population Data:

	Country	Continent	Population_2022	Area(km²)	Density(per km²)	Growth_Rate	World_Population_Percentage
0	Afghanistan	Asia	41128771	652230	63.0587	1.0257	0.52
1	Albania	Europe	2842321	28748	98.8702	0.9957	0.04
2	Algeria	Africa	44903225	2381741	18.8531	1.0164	0.56
3	American Samoa	Oceania	44273	199	222.4774	0.9831	0.00
4	Andorra	Europe	79824	468	170.5641	1.0100	0.00
5	Angola	Africa	35588987	1246700	28.5466	1.0315	0.45
6	Anguilla	North America	15857	91	174.2527	1.0066	0.00
7	Antigua and Barbuda	North America	93763	442	212.1335	1.0058	0.00
8	Argentina	South America	45510318	2780400	16.3683	1.0052	0.57
9	Armenia	Asia	2780469	29743	93.4831	0.9962	0.03

Figure 1: First 10 Rows of World Population Data

First 10 Rows of Air Pollution Data:

	Country	AQI_Value	AQI_Category	CO_AQI_Value	CO_AQI_Category	...	Ozone_AQI_Category	NO2_AQI_Value	NO2_AQI_Category	PM25_AQI_Value	PM25_AQI_Category
0	Russian Federation	51	Moderate	1	Good	...	Good	0	Good	51	Moderate
1	Brazil	41	Good	1	Good	...	Good	1	Good	41	Good
2	Italy	66	Moderate	1	Good	...	Good	2	Good	66	Moderate
3	Poland	34	Good	1	Good	...	Good	0	Good	20	Good
4	France	22	Good	0	Good	...	Good	0	Good	6	Good
5	United States of America	54	Moderate	1	Good	...	Good	11	Good	54	Moderate
6	Germany	62	Moderate	1	Good	...	Good	3	Good	62	Moderate
7	Belgium	64	Moderate	1	Good	...	Good	7	Good	64	Moderate
8	Russian Federation	54	Moderate	1	Good	...	Good	1	Good	54	Moderate
9	Egypt	142	Unhealthy for Sensitive Groups	3	Good	...	Moderate	9	Good	142	Unhealthy for Sensitive Groups

Figure 2: First 10 Rows of Air Pollution Data

5.2 Data Format

Reason for Choosing SQLite:

- **Efficiency:** Efficient storage and querying.
- **Portability:** Easily transferable and shareable.
- **Ease of Use:** Simple to use with pandas for data manipulation.

5.3 Critical Reflection and Potential Issues

Strengths: Clean and consistent data, efficient querying capabilities.

Potential Issues:

- **Data Completeness:** Dropping rows with missing values may lead to the loss of valuable information.
- **Aggregation Assumptions:** Aggregating air pollution data assumes that mean values are representative, which may overlook local variations.
- **Geographical and Temporal Coverage:** Data is limited to certain regions or periods which may affect trend analysis.
- **Analysis Limitations:** Simplifying data by aggregation might oversimplify real-world dynamics.