Data Report: Exploring the Relationship Between Air Quality and Respiratory Health

Question

The main question driving this project is: **How does air quality impact respiratory health across different regions in California, and what trends can be observed over time?** This study explores the relationship between pollution levels (PM2.5) and respiratory conditions, leveraging publicly available datasets.

Data Sources

Air Quality Data

Source: EPA Air Quality System (AQS) **Reason for Selection:** Provides detailed PM2.5 pollution measurements across California, critical for understanding environmental health impacts. **Content:** Includes pollutant levels, sampling locations, dates, and measurement units. **Structure and Quality:**

- Structured tabular data with attributes such as `state_code`, `date_local`, `sample_measurement`.
- Missing values handled during preprocessing.
- High reliability as data is maintained by the EPA.

License and Obligations: Licensed under open access by the EPA. Obligations include attribution, addressed in the project documentation. License Details

Respiratory Health Data

Source: CDC Chronic Disease Indicators Dataset **Reason for Selection:** Provides health metrics for respiratory conditions (e.g., asthma) at state and county levels. **Content:** Includes respiratory health indicators, demographic stratifications, and annual statistics. **Structure and Quality:**

- Structured tabular data with fields such as `topic`, `state_code`, `respiratory_value`, and `year`.
- Data contains noise, requiring filtering for relevant indicators.

License and Obligations: Open Data license with attribution requirements, fulfilled via documentation. License Details

Data Pipeline

Overview: The pipeline automates data fetching, cleaning, transformation, merging, and output, consisting of the following stages:

1. **Extraction:** Fetches air quality data via the EPA API and respiratory data from the CDC dataset, saved as raw CSV files.
2. **Transformation:**
   - **Air Quality Data:** Handled missing/negative values, converted date formats, and replaced numeric state codes with state abbreviations.
   - **Respiratory Data:** Filtered for relevant topics (e.g., asthma), renamed columns for consistency, and standardized date formats.
3. **Loading and Analysis:** Merges datasets using `state_code` and `year` as keys and outputs a merged CSV file for further analysis.

Pipeline Diagram:

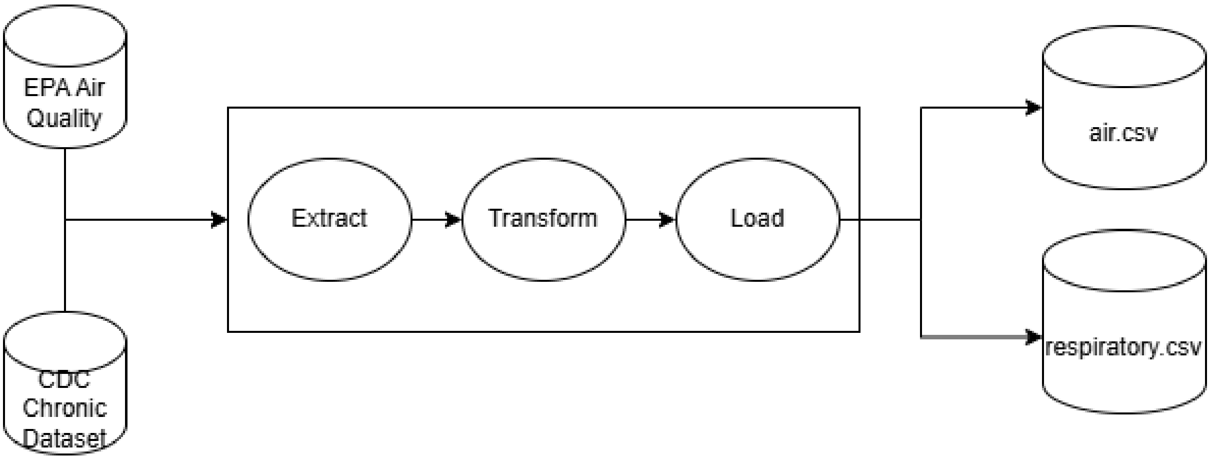

Figure 1: Data Pipeline Diagram

Technology:

- **Language:** Python
- **Libraries:** `pandas`, `numpy`, `requests`, `threading`

Issues and Resolutions:

- **Inconsistent column names across datasets:** Renamed columns to align (`state_code`, `year`).
- **Numeric state codes in air quality data:** Converted to abbreviations (e.g., `6` to `CA`).
- **Mismatched keys during merging:** Standardized key columns before merging.

Meta-quality Measures:

- Exceptions raised for missing or malformed data, with logs for progress.
- Dynamic column validation ensures compatibility with changes in data structure.

Result and Limitations

Output Data: The pipeline produces a merged dataset (`merged_data.csv`) containing air quality metrics and respiratory health data for California.

Data Quality:

- Cleaned and structured for analysis.
- Limitations include missing historical data for specific counties and aggregation obscuring localized trends.

Output Format:

- **Format:** CSV for compatibility with analysis tools.
- **Reason:** Simple, universal format for structured data.

Critical Reflection:

- While the datasets provide a solid basis for analysis, their granularity limits deeper insights into regional disparities.
- Incorporating additional datasets, such as socioeconomic indicators, could enhance the depth of analysis.