Data Analysis and Pipeline Report

Kateryna Kolesova

1. Question

What are the relationships between air quality levels and chronic disease prevalence in New York, year 2022?

This project aims to identify potential correlations or trends between air pollution and the prevalence of chronic diseases.

2. Data Sources

Data Source 1: Air Quality in New York

- **Description**: Dataset contains information on New York City air quality surveillance data.
- **Source**: Link to the website. Data provided by Department of Health and Mental Hygiene (DOHMH). Owner of the data set is NYC Open Data.
- **Data set dictionary**
- **License**: Open Data Commons Public Domain Dedication and License (PDDL). Laws and Terms of Use.

Data Source 2: Chronic Disease Data in New York

- **Description**: This dataset includes the prevalence rates of chronic diseases (e.g., asthma, cardiovascular diseases) in the USA, which we later narrowed down to New York.
- **Source**: Link to the website.
- **License**: Open Database License (ODbL)

Data Exploration Overview

Structure and Quality

Chronic Disease Dataset

Column	Sample Value
YearStart	2008
YearEnd	2012
LocationDesc	New York
Topic	Cancer

Column	Sample Value
`Question`	Cancer of the oral cavity and pharynx, mortality
`DataValue`	329
`DataValueUnit`	Average Annual Number
`GeoLocation`	POINT (-75.54, 42.83)

Insights

- Covers various health topics, including mortality rates related to diseases like cancer.
- Data includes geographic coordinates (`GeoLocation`) and temporal ranges (`YearStart`, `YearEnd`).

Air Quality Dataset

Column	Sample Value
`name`	Nitrogen dioxide (NO2)
`measure`	Mean
`geo_place_name`	Pelham - Throgs Neck
`time_period`	Summer 2022
`start_date`	2022-06-01
`data_value`	12.0
`measure_info`	ppb

Insights

- Tracks pollutants such as NO_2, PM2.5, and O_3.
- Includes detailed geographic (`geo_place_name`) and temporal (`start_date`) information.
- Data values are standardized in units like `ppb` (parts per billion) and `mcg/m³`.

Next Steps

1. Check for overlapping locations between:
    - `LocationDesc` (Chronic Disease dataset)
    - `geo_place_name` (Air Quality dataset)
2. Perform exploratory data analysis to clean and align:
    - **Temporal data** (e.g., yearly vs. seasonal/daily reporting)
    - **Geographic data** (e.g., standardizing location names)
3. Assess missing or inconsistent values in each dataset.

---

3. Data Pipeline

Overview

This Python pipeline:

1. Downloads datasets from provided URLs.
2. Saves them as CSV files in a specified directory.
3. Filters the data for a specific location (`New York`) or time period (e.g., 2022).
4. Stores the filtered data in SQLite databases for easier querying and analysis.

Key Steps

1. Download Data

- Uses `requests` to fetch datasets from given URLs.

2. Save Data Locally

- Saves datasets as CSV files in the specified directory (`../data`).

3. Filter Data

- Applies filters for:
  - **Location**: Using the `LocationDesc` column to extract rows containing "New York".
  - **Time Period**: Matches specific years or ranges for columns `Time Period`, `YearStart`, `YearEnd`.

4. Save to SQLite

- Converts filtered datasets into SQLite tables.

---

4. Result and Limitations

Output Data

- **Structure**:
  - The final dataset includes columns for date, location, air pollutant levels, and chronic disease prevalence rates.
  - Data format: CSV and Parquet for scalability.
- **Quality**:
  - Geographic data successfully aligned between datasets.

Limitations

- The `filter_time` logic has issues with checking the time columns and should be fixed.
- **Temporal Coverage**: Air quality data is daily, while chronic disease data is aggregated annually, which limits temporal precision in analysis.
- **Causation vs. Correlation**: The project identifies correlations but cannot confirm causal relationships due to external confounding factors (e.g., socioeconomic status, healthcare access).
