data-report

November 28, 2024

1 Data Report

1.1 Question

What insights can be gained from the analysis of chronic disease indicators and their correlation with the leading causes of death in the United States?

1.2 Data Sources

1.2.1 1. Chronic Disease Indicators Dataset

- **Why Chosen**: This dataset provides detailed information about chronic diseases, including prevalence and public health impact.
- **Source**: Chronic Dataset
- **Data Type**: CSV
- **Structure**:
  - Columns include:
    - `YearStart`: Start year of data collection.
    - `LocationDesc`: State or territory name.
    - `Topic`: Chronic disease topic (e.g., Cancer, Diabetes, etc).
    - `Question`: A specific question about the topic (e.g., "Mortality from diabetes, etc").
    - `DataValueType`: Type of value (e.g., "Age-adjusted Rate, Number, etc").
    - `DataValue`: Numerical value (e.g., mortality rate, etc).
    - `Stratification1`: Demographic breakdown (e.g., "Overall,"Gender, etc").
- **License**:
  - Licensed under the Open Database License (ODbL). Fulfill obligations by proper attribution and sharing derivatives under the same license. License Info

1.2.2 2. NCHS Leading Causes of Death Dataset

- **Why Chosen**: Essential for understanding mortality trends and their correlation with chronic diseases.
- **Source**: NCHS Dataset
  - `Cause Name`: Cause of death (e.g., "Heart Disease").
  - `State`: State or territory name.
  - `Deaths`: Number of deaths recorded.
  - `Age-adjusted Death Rate`: Standardized mortality rate per 100,000 population.
- **License**:
  - Public domain, with attribution recommended as a courtesy. License Info

1.3 Data Pipeline

1. **Download**: The pipeline downloads datasets programmatically using Python's `requests` library to ensure we always work with the most recent data.
2. **Filtering and Cleaning**:
   - For chronic disease data:
     - Filter `DataValueType == "Age-adjusted Rate"`.
     - Keep only rows with `Demographic == "Overall"` (The "Overall" category represents aggregated data for the entire population and the NCHS dataset does not provide granular demographic data and I am looking for insighs on the whole population).
     - Focus on questions containing "Mortality."
   - For mortality data:
     - Here I retain only relevant causes of death based on predefined mappings. The mapping aligns broad chronic disease topics (e.g., "Diabetes") with specific causes of death (e.g., "Diabetes-related deaths") to enable meaningful merging and analysis of the two datasets.
3. **Merging**:
   - Merge chronic disease data with mortality data on `Year`, `State`, and `Cause_Name`.
4. **Output**:
   - Store the final cleaned dataset in SQLite and CSV formats.

1.3.1 Technologies Used

- **Programming Language**: Python
- **Libraries**: `pandas`, `requests`, `sqlite3`
- **Storage**: SQLite for structured queries, CSV for sharing and analysis.

1.4 Challenges in the Data Pipeline

1.4.1 Multiple `DataValueType` Rows

The Chronic dataset included two rows per `Year`, `State`, and `Topic` due to: - **Number**: Raw count. - **Age-adjusted Rate**: Standardized per 100,000 population.

Solution: Retained only `Age-adjusted Rate` for consistency and to simplify merging with the NCHS dataset.

1.4.2 Mismatch Between Topics and Cause Names

The `Topic` column in the Chronic dataset did not directly align with the `Cause Name` column in the NCHS dataset.

Solution: Created a mapping dictionary to align `Topic` with `Cause Name` (e.g., "Cardiovascular Disease" → "Heart Disease"). Topics without matches were excluded.

1.4.3 Challenges in Mapping

1. **Unmatched Topics**: Some Chronic topics lacked corresponding NCHS causes.
   Solution: Included only mappable topics.

2. **Ambiguous Terminology**: Terms like "Respiratory Diseases" could match multiple causes. **Solution**: Used domain knowledge for precise mappings.

1.5 Data Structure

The final merged dataset contains 832 rows and 6 columns: `Year` (the year the data was collected), `State` (the state or territory where the data was recorded), `Topic` (e.g., "Cancer," "Diabetes"), `Deaths` (total deaths recorded in the NCHS dataset), `Age_Adjusted_Death_Rate` (the standardized death rate per 100,000 population), and `ChronicDiseaseValue` (the age-adjusted rate of the chronic disease indicator).

1.6 Data Quality

- **Completeness**: Selected columns are complete; rows with missing or unmatched values were excluded.
- **Accuracy**: High reliability due to CDC-sourced data.
- **Consistency**: Uniform column formats and retained only `Age-adjusted Rate` rows for standardization.
- **Relevance**: Focused on mortality-related indicators for chronic diseases.
- **Data Format**: SQL for structured analysis and CSV for sharing.

1.7 Limitations

1.7.1 Data Reduction Summary

The Chronic dataset originally had over 1.1 million rows, and the NCHS dataset contained approximately 10,000 rows. After processing, the final merged dataset was reduced to 832 rows, focusing only on the topics "Chronic Obstructive Pulmonary Disease" and "Diabetes."

1.7.2 Reasons for Reduction

Filtering retained only rows with `DataValueType == "Age-adjusted Rate"`, `Demographic == "Overall"`, and questions containing "Mortality." Merging used an inner join on `Year`, `State`, and `Cause_Name`, keeping only matching rows. Topics were limited due to predefined mapping and lack of corresponding data in NCHS.

1.7.3 Implications

This reduction caused a loss of detail by excluding many rows and topics like "Cardiovascular Disease." The focus on "Chronic Obstructive Pulmonary Disease" and "Diabetes" limits scope, and the small sample size (832 rows) may impact the robustness of the analysis.