1 Questions

- Are there correlations between chronic disease indicators (e.g., diabetes prevalence) and specific death rates (e.g., due to heart disease)?
- What were the leading causes of death each year between 2020 and 2023?

2 Data Sources

2.1 Descriptions of Data Sources

- **Datasource 1: Monthly Provisional Counts of Deaths by Select Causes (2020–2023)**

  The dataset includes the total counts of deaths attributed to different causes like heart disease, diabetes, cancer and COVID-19. It provides data spanning both time-related (monthly from 2020 to 2023) and geographic coverage for the United States. [1]

- **Datasource 2: U.S. Chronic Disease Indicators (CDI)**

  The dataset includes information on chronic disease indicators highlighting prevalence rates, demographic distributions and trends over time. By providing the insights of underlying health conditions that are leading to death, this dataset helps in getting a deeper understanding about the factors contributing to deaths. [3]

2.2 Structure and Quality of Data Sources

- **Monthly Provisional Counts of Deaths by Select Causes**

  The dataset represents the monthly provisional counts of deaths characterized by cause. Each columns represents distinct types of disease associated with the deaths, while the rows display the total death counts with the corresponding dates, months, years and different causes of death (e.g., heart disease, diabetes, cancer). There are a few missing values, but most columns have no missing data. Some columns are empty, however, they are not necessary for the analysis

	Data As Of	Start Date	End Date	Jurisdiction of Occurrence	Year	...	flag_accid	flag_mva	flag_suic	flag_homic	flag_drugod
0	09/27/2023	01/01/2020	01/31/2020	United States	2020	...	NaN	NaN	NaN	NaN	NaN
1	09/27/2023	02/01/2020	02/29/2020	United States	2020	...	NaN	NaN	NaN	NaN	NaN
2	09/27/2023	03/01/2020	03/31/2020	United States	2020	...	NaN	NaN	NaN	NaN	NaN
3	09/27/2023	04/01/2020	04/30/2020	United States	2020	...	NaN	NaN	NaN	NaN	NaN
4	09/27/2023	05/01/2020	05/31/2020	United States	2020	...	NaN	NaN	NaN	NaN	NaN

Figure 1: First 5 rows of Monthly Provisional Counts of Deaths by Select Causes.

- **U.S. Chronic Disease Indicators (CDI)**

  The dataset is represented in tabular format where the rows represents health indicators and columns are grouped by demographics (e.g., gender, race, geographic location). The primary columns in the dataset are year, location, topic (indicator), questions and datavalue. Some columns are empty, however, they are not necessary for the analysis.

	YearStart	YearEnd	LocationAbbr	...	StratificationID2	StratificationCategoryID3	StratificationID3
0	2010	2010	OR	...	NaN	NaN	NaN
1	2019	2019	AZ	...	NaN	NaN	NaN
2	2019	2019	OH	...	NaN	NaN	NaN
3	2019	2019	US	...	NaN	NaN	NaN
4	2015	2015	VI	...	NaN	NaN	NaN

Figure 2: First 5 rows of U.S. Chronic Disease Indicators (CDI).

2.3 Licenses and Permissions

- **Monthly Provisional Counts of Deaths by Select Causes (2020–2023)**

  The data is licensed under the `U.S. Government Work (Public Domain)`. The dataset is publicly available and produce by the U.S. Government. The dataset is free to use, with the only requirement being tp provide proper attribution to the U.S. government. Additionally, to ensure that no logos or other content are used in a way that implies endorsement by the U.S. government. [4]

- **U.S. Chronic Disease Indicators (CDI)**

  This dataset is available for free use under`Open Database License (ODbL)`. User are required to provide proper attribution to the original source. Additionally, any derived work must be shared under the same license. [2]

2.4 Data Pipeline

The data pipeline is developed using Python. It contains three main modules: extract, transform, load. These modules contain the following functionalities.

We start by reading the metadata from a JSON file (`datasources.json`) to gather information about the data sources like URL, and columns to be removed. The process starts by extracting the data from the URL using the `CsvExtractor` function, which retrieves the dataset in a CSV. Next, unnecessary columns are removed using the `DeleteColumns` function, based on the configuration provided in the metadata. If a filtering query is specified for the dataset, the `FilterRows` function is used to filter the data accordingly. Then, empty values in the dataset are filled using the `FillEmptyValues` function to ensure completeness. Finally, the transformed dataset is loaded into the SQLite database (`ChronicHealthTrends.db`) using the `LoadDfToSqlite` function, storing the data with the appropriate dataset name.

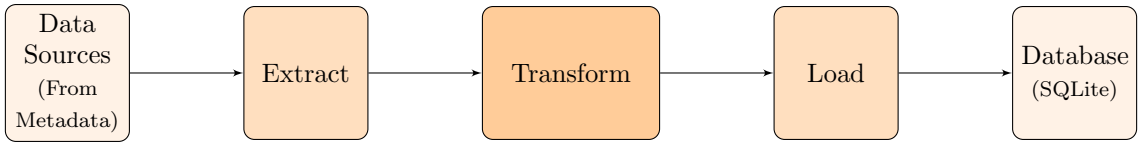

Figure 3: ETL Process Diagram

3 Results and Limitations

The resulting datesets are stored in SQLLite in tabular format as it is the most efficient and reliable way to store collective data. The pipeline can maintain the quality of data. The resultant datasets

- fulfill the requirement that is needed to answer the questions.
- time domain is overlapping and fitting
- ensure format is consistent

The datasets on chronic disease indicators and mortality statistics can be compared and analyzed to uncover correlations between chronic health conditions and mortality trends across various demographic and geographic segments. This analysis provides a detailed understanding of the underlying health factors contributing to death rates. However, a limitation exists regarding the granularity of chronic disease data, as it may not provide sufficient specificity directly link certain health indicators to the cause of mortality.

References

[1] Monthly provisional counts of deaths by select causes, 2020-2023. https://catalog.data.gov/dataset/monthly-counts-of-deaths-by-select-causes-2020-2021-2785a.

[2] Open database license (odbl). https://opendefinition.org/licenses/odc-odbl/.

[3] U.s. chronic disease indicators (cdi), 2023. https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi.

[4] U.s. government work (public domain). https://www.usa.gov/government-copyright.