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 number of deaths happened due to different causes like heart disease, diabetes, cancer and COVID-19. It provides data spanning over both time-related (monthly from 2020 to 2023) and geographic coverage for United States. [1]

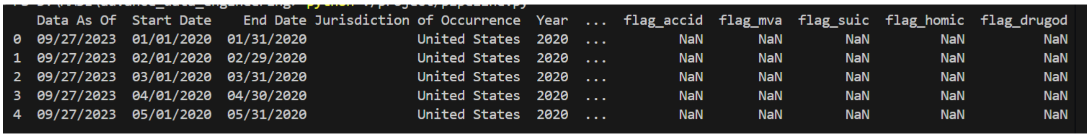
- **Datasource 2: U.S. Chronic Disease Indicators (CDI)**

  The dataset provides information of chronic disease indicators including 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 the deeper understanding factors contributing to deaths. [3]

2.2 Structure and Quality of Data Sources

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

  The dataset is structured in a way that the causes of deaths are in columns and rows represent the months, well structured for time series analysis. Key columns include Year, Month and different causes of death (e.g., heart disease, diabetes, cancer). Null values are not that much and most columns are completely populated. Some columns are empty but they are not needed for the analysis.

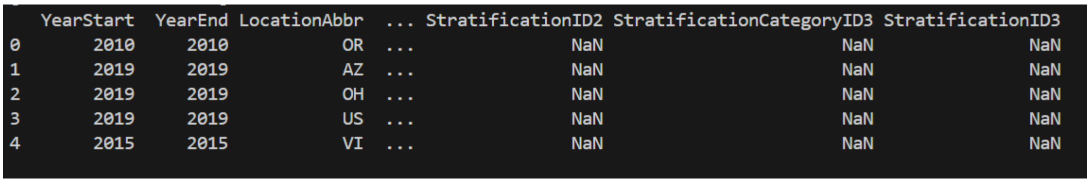


```
   Data As Of  Start Date    End Date Jurisdiction of Occurrence  Year  ...  flag_accid  flag_mva  flag_suic  flag_homic  flag_drugod
0  09/27/2023  01/01/2020  01/31/2020              United States  2020  ...         NaN       NaN        NaN         NaN          NaN
1  09/27/2023  02/01/2020  02/29/2020              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 were rows represent health indicators and columns is grouped by demographics (e.g., gender, race, geographic location). Key columns are Year, location, topic (is the indicator), questions and Datavalue. Some columns are empty but they are not needed 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 license under the `U.S. Government Work (Public Domain)`. The dataset is publicly available and produce by US Government. I can freely use this dataset to follow there restriction I will provide proper attribution to the U.S. government as the source of the data, While ensuring that I will not use any logo or anything that implies that it is endorse by the U.S. government. [4]

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

  This dataset is licensed under `Open Database License (ODbL)`. I am allow to free use this but need to provide proper attribution. I will adhere to this restriction by properly citing this dataset in my report and analysis. [2]

2.4 Data Pipeline

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

The pipeline begins by reading the metadata from a JSON file (`datasources.json`) to gather information about the data sources like URL, and columns to remove. The process starts by extracting the data from the URL using the `CsvExtractor` function, which retrieves the dataset in CSV format. 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. Now, empty values in the dataset are filled using the `FillEmptyValues` function to ensure completeness. Finally, the transformed dataset is loaded into a SQLite database (`ChronicHealthTrends.db`) using the `LoadDfToSqlite` function, storing the data with the appropriate dataset name.

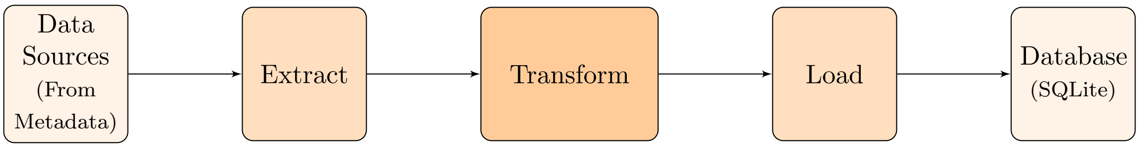


Figure 3: ETL Process Diagram

3 Results and Limitations

Output result of the data pipeline is stored in SQLLite in table format as this 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
- 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 enables a comprehensive understanding of the health factors contributing to death rates. However, a limitation arises with the granularity of chronic disease data, as it may not provide sufficient specificity to link certain health indicators directly to mortality causes.

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.