Q - A brief description of your data and your reason(s) for choosing your

dataset(s).

Ans - I chose this dataset because it provided health related and demographic data across 3,000+ cities in the U.S. Each row included information about each city such as the number of hospitals and clinics, population, income, and health indicators varied from life expectancy, uninsured, obesity, diabetes, and other health indicators. This dataset provides a broad spectrum of public health dimensions across health hotspots and gives opportunities to identify patterns and differences.

Why I Selected This Dataset:

* Multiple aspects of health: This dataset had multiple health-related content areas son it provided many opportunities to explore how at least two aspects of health are related to three HLT variables at the same time (i.e., access, income, health).
* Large and Complete Dataset: It incorporated thousands of entries and at least 10 columns related to my topic so there was ample data to analyze and apply conclusions on.
* Theme-based regional comparisons: Because it incorporates cities in all 50 states, the datasets provided unformed comparisons to each city, health, demographic and population trends (similar to theme) to highlight regional differences.
* Opportunities to explore concepts or relationships: dataset compared you could look directionally across the counties and explore interesting relationships between income or costs associated with access to hospitals to life expectancy.

Q - A description of your data collection methods/sources.

Ans - The data for this project came from a compiled CSV file that combined health care and demographic data for all cities from each state in the United States. Although the dataset was curated, there appears to be data available from publicly available government databases or health care databases such as:

* U.S. Census Bureau - population and income statistics
* Centers for Disease Control and Prevention - health indicators such as obesity, diabetes, and life expectancy
* Centers for Medicare & Medicaid Services - uninsured rates and access to health care
* State health departments - counts of hospitals and clinics from where the hospitals and clinics are located.

These sources are used frequently in public health research and contain reliable and up to date data. The dataset was constructed into a tidy table with consistent column names, and thus, the dataset was ready to analyze with little preprocessing.

Q - A paragraph describing your data cleaning/wrangling process.

Ans - To prepare the dataset for analysis, a series of cleaning and transformation steps were performed using OpenRefine. The primary goals were to remove irrelevant information, handle missing values, and simplify the dataset for more focused analysis. The steps included:

1. Column Removal: Multiple columns that were either metadata, redundant, or not relevant to the analytical goals were removed. These included geographic codes (geo\_fips, state\_fips), technical fields (geo\_level, census\_parent\_shape\_year), and documentation-related fields such as disclaimers and suggested citations. This reduced noise and helped streamline the dataset for analysis.
2. Standardizing Missing Data: Several columns (num, denom, lci, uci) had placeholder values of -999 to indicate missing data. These were replaced with NA to standardize missing value representation, making the dataset cleaner and more compatible with statistical tools.
3. Data Filtering: Only rows where the group\_name was "Total" and the metric\_name matched a specific set of public health indicators (e.g., Obesity, Diabetes, Firearm Suicides) were retained. This focused the dataset on relevant metrics for analysis across the total population, avoiding subgroup data that could complicate comparisons.
4. Final Cleanup: After transformation, additional columns (num, denom, data\_indicator) were removed as they don’t have a good amount of non-null values.

Overall, the dataset was extensively cleaned and is now in a structured, analysis-ready format with clearly defined health indicators across U.S. cities. No additional wrangling is necessary at this stage.

Q - 10 questions you formulated and the corresponding answers/charts/visuals to your own questions about your data, including machine learning algorithms or statistical analysis methods you applied to get your answers, and what software(s) you used.

Ans - **Health & Outcome Focused Questions**

1. Which cities have the highest and lowest rates of obesity?

A graph of a number of cities

Description automatically generated

A graph with green lines

Description automatically generated

**Interpretation**: There is clear geographic disparity in obesity rates, with specific cities consistently ranking at the extremes. This may suggest regional lifestyle, access to healthy food, or socioeconomic status as contributing factors.

1. Is there a correlation between frequent mental distress and chronic physical health issues like diabetes or cardiovascular disease?

A red and blue squares with white text

Description automatically generated

**Interpretation**: A moderate to strong positive correlation likely exists. This supports the idea that mental health is intertwined with physical health, particularly chronic conditions.

1. Do cities with higher rates of firearm suicides also have higher rates of mental distress or binge drinking?

A diagram of a heatmap

Description automatically generated

**Interpretation**: A visible upward trend would imply a statistically meaningful link between mental distress/binge drinking and firearm suicide rates. This supports public health initiatives focusing on mental health to reduce suicides.

**Mental Health & Substance Use Questions**

1. What is the geographic distribution of frequent mental distress across states?

**A graph of mental distress

Description automatically generated**

**Interpretation**: Certain states or regions (possibly in the South or Midwest) show higher mental distress rates, suggesting potential need for targeted mental health services.

**Geographic Comparisons**

1. Are certain states or regions consistently above or below national averages in preventable death metrics?

A graph of a number of columns

Description automatically generated with medium confidence

**Interpretation**: Some states consistently underperform in preventable health outcomes. This reflects systemic public health issues, including access to care, socioeconomic inequality, or public policy gaps.

1. Do cities with poor access to dental care also report higher physical distress or chronic disease rates?

A graph showing a red line between blue dots

Description automatically generated

**Interpretation**: A negative correlation is likely seen, suggesting a significant relationship. This emphasizes the role of dental health as a predictor or marker for broader health outcomes.

**Source Reliability or Methodological Questions**

1. Do health estimates differ significantly between data sources for similar indicators, and why might that be?

A graph with numbers and a bar chart

Description automatically generated with medium confidence

**Interpretation**: The visuals indicate that data on premature deaths from all causes is primarily sourced from NJSHAD and NVSS MCDD, with other sources contributing only marginally. This suggests a strong reliance on these two primary data providers for mortality metrics.

**Preventive Care & Mortality Questions**

1. Is there a significant association between lack of dental care and chronic diseases like diabetes?

A graph of blue dots

Description automatically generated

**Interpretation**: A negative relationship between poor dental care and diabetes prevalence would affirm the systemic connection between oral and overall health.

1. What cities have the highest rates of premature deaths from all causes?

A graph showing the number of cities

Description automatically generated

**Interpretation**: Certain cities(Decatur) exhibit significantly higher rates, possibly linked to socioeconomic status, healthcare access, or environmental factors.

**Demographic Disparity Questions**

1. Are there health disparities between cities in the same state?

A chart of different colored squares

Description automatically generated

**Interpretation**: Significant disparities were likely observed, even within the same state. This reflects localized policy effectiveness, infrastructure, or community health resources.

Q - A separate page for your data dictionary.

Ans - metric\_name:

The specific health or demographic indicator being measured (e.g., Obesity, Diabetes, Binge Drinking). This column defines the type of data each row represents.

data\_period:

The year or range of years during which the data was collected or reported (e.g., 2020, 2020–2021). This allows for temporal analysis and trend comparisons.

est (Estimate):

The main numerical value or estimate of the selected metric for a given location and time period. For example, this might be the obesity rate or the percentage of people with high blood pressure.

lci (Lower Confidence Interval):

The lower bound of the 95% confidence interval for the estimate, indicating the range within which the true value is likely to fall.

uci (Upper Confidence Interval):

The upper bound of the 95% confidence interval for the estimate. Together with lci, it reflects the uncertainty around the est value.

geo\_name:

The name of the geographic location (typically a city or county) for which the metric is reported.

state\_abbr:

The two-letter abbreviation of the U.S. state in which the location (geo\_name) is found.

period\_type:

Indicates whether the data covers a single year or a multi-year period (e.g., Single Year, 3-Year Average), which affects how data should be interpreted.

source\_name:

The name of the original data source or agency that provided the data (e.g., CDC, U.S. Census Bureau). This is useful for verifying data quality and methodology.

Q - Your citation/sources such as links, etc.

Ans - Citations / Data Sources

1. **City Health Dashboard**

Main source of the integrated dataset, offering localized health metrics for U.S. cities.

Website: https://www.cityhealthdashboard.com

1. **Python**

General-purpose programming language used for data analysis and visualization.

Website: https://www.python.org

1. **Pandas**

Open-source data manipulation and analysis library for Python, ideal for working with tabular data.

Website: https://pandas.pydata.org

1. **Matplotlib**

Comprehensive library for creating static, animated, and interactive visualizations in Python.

Website: https://matplotlib.org

1. **Seaborn**

Statistical data visualization library built on top of Matplotlib, offering high-level plotting functions.

Website: https://seaborn.pydata.org