CPSC 368 Research Paper

KNM Neighbours (Nicholas Tam (45695970), Kevin Shiao (73239121), Minghao Wang (56536469))

Abstract

This study investigates the impact of health insurance rates from three perspectives: sex, state, and disease type. Datasets were drawn from the U.S. Chronic Disease Indicators (Centers for Disease Control and Prevention) and the Kaiser Family Foundation (KFF). Results show that male CHD mortality rates per 100,000 individuals exceed those of females in both Massachusetts (36 vs. 16 deaths, respectively) and Texas (57 vs. 29 deaths), a pattern aligning with existing literature highlighting higher CHD risks in men. Since males have a lower uninsured rate compared to females, this implies that the uninsured rate could be a factor that influences the average mortality rate. Although Texas exhibits a notably higher uninsured rate than Massachusetts (24.5% vs. 4.4%), overall CHD and cancer mortality outcomes are relatively similar, suggesting that the uninsured rate alone may not fully explain mortality differences. Cancer analyses reveal lung cancer as having the highest mortality rate among examined types (breast, cervical, colorectal, lung, and prostate) in both states. However, Massachusetts generally shows higher rates, except for cervical cancer. State-level analyses using Support Vector Regression (SVR) indicate that, for younger age groups (0–44, 45–64), uninsured rates can predict variations in CHD mortality, whereas for older adults (65+), geographic location exerts greater influence. These findings establish a foundation for continued research in health insurance and broader public health disciplines. The majority of our associated code and data is provided in the cpsc368_knm_project Github repository (Tam, Shiao, Wang, 2025).

1. Introduction

Modern technologies have improved our quality of life, making everything more convenient than ever by reducing workload and stress from various sources. This allows people to focus on their health more efficiently. With modern health information technologies, users can receive immediate feedback on their physical condition anytime, anywhere (Li et al., 2019). Health insurance is essential and beneficial, as uninsured individuals often experience poorer health and receive less medical care, often with delays (Bovbjerg, Hadley, 2006).

However, the extent to which health insurance impacts health remains debatable. As Levy and Meltzer suggest, determining whether health insurance significantly influences health will require substantial investment in social experiments (Levy, Meltzer, 2008). According to the Institute of Medicine (US) Committee on the Consequences of Uninsurance, there have been various studies and examinations across the past several decades on both the relationship between health insurance and health outcomes, and the mechanisms used to measure and determine that relationship (Institute of Medicine (US) Committee on the Consequences of Uninsurance, 1970). Across a body of studies that use a variety of data sources and different analytic approaches, researchers have determined a consistent, positive relationship between health insurance coverage and health-related outcomes, with the best evidence suggesting that health insurance is associated with more appropriate use of health care services and better health outcomes for adults (Institute of Medicine (US) Committee on the Consequences of Uninsurance, 1970a).

Although this paper does not directly answer the ultimate question or fill the gap, it contributes by presenting significant findings that serve as supporting evidence, aiming to attract attention in the healthcare and health insurance fields. Specifically, we explore how health insurance coverage impacts health outcomes among U.S. adults. The impact of health insurance will be measured in three ways: (1) by sex (male and female), focusing on coronary heart disease mortality by sex; (2) by state, examining coronary heart disease mortality across different states; and (3) by disease, comparing coronary heart disease mortality with various cancer mortalities.

2. Methodology

The datasets 'U.S. Chronic Disease Indicators' and 'Health Insurance Coverage of Adults Ages 19-64' are from HealthData.gov and KFF, respectively. The 'U.S. Chronic Disease Indicators'' dataset contains 4820 observations, and the attributes used are YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, and Stratification1 (Centers for Disease Control and Prevention, 2024). On the other hand, the datasets 'Health Insurance Coverage of Adults Ages 19-64', 'Health Insurance Coverage of Women Ages 19-64', and 'Health Insurance Coverage of Men Ages 19-64' from KFF each contain 52 observations, one for each state in the US, and the attributes used are Location and Uninsured (KFF, 2024). Data exploration, evaluation, and cleaning will be conducted using Python. To more effectively explore the datasets, the two tables will be joined by state. Once these processes are completed, the cleaned data will be structured and stored in an SQL database, with key questions addressed using SQL queries.

Since the combined table does not include population percentages for males and females, additional methods are needed to approximate insurance distribution by sex. To measure the impact by sex, total population data and male/female population percentages for the examined states will be obtained from government-authorized websites to ensure accuracy and reliability. From there, the percentage split between males and females in the examined states is determined by:

$$\%_{sex} * N_{total} = N_{sex}$$

$$PI_{x} * N_{total} = N_{x}$$

$$PI_{sex_{x}} = PI_{x} * (\frac{N_{sex}}{N_{total}})$$

Where $\%_{sex}$ is the population percentage of sex (male or female), N_{total} is the total population, N_{sex} is the population of sex (male or female), PI_{x} is the proportion insured by x (insured or uninsured), N_{x} is the proportion of the population insured by x (insured or uninsured), and $PI_{sex_{x}}$ is the proportion of sex (male or female) insured by x (insured or uninsured).

2.1 Data Evaluation

Before each evaluation, QQ plots and residual plots will be created to validate model assumptions and ensure robustness and accuracy. Specifically, we will check the normality of residuals, homoscedasticity, and independence. Appropriate corrective measures, such as data transformations, will be applied if any assumptions are violated

Impact by Sex

For this analysis, we intended to use Multiple Linear Regression (MLR) to examine the relationship between uninsured rates and coronary heart disease (CHD) mortality rates across males and females in Texas and Massachusetts. The model would have included uninsured rates as a continuous predictor and sex as a categorical variable. Additionally, we would have incorporated an interaction term between the uninsured rate and sex. This would have allowed us to determine if higher uninsured rates are more strongly associated with CHD mortality in one sex compared to the other. By employing MLR with interaction, we allow for a deeper understanding of relationships between variables by capturing the potential effects between variables, making our approach more scientifically robust. For more effective interpretations, we would also visualize our model to display how the relationship between uninsured rates and CHD mortality varies by sex.

Upon further review of how our datasets and questions have been organized and cleaned, we found that the planned data evaluation may be unnecessary and infeasible. Since we selected only two states to analyze the impact of sex, we do not have enough data points to conduct Multiple Linear Regression. As a result, our data evaluation will have to be done indirectly unless we use all available data points.

Impact by State

We will also investigate the relationship between uninsured rates and coronary heart disease (CHD) mortality across all US states. First, we will gather data on each state's uninsured and CHD mortality rates. Then, we will use Support Vector Regression (SVR) to analyze mortality rates and uninsured rates across the US to see any notable relationships between them. We chose to apply SVR because it can capture the non-linear trends in the data that other models may not be able to capture. Additionally, we will perform hyperparameter optimization to fine-tune our model and ensure it is as accurate as possible.

Impact by Disease

Lastly, to compare the impact health insurance has on different diseases, we will develop two Poisson regression models: one to investigate the relationship between uninsured rates and coronary heart disease mortality rates and another to examine the same relationship for cancer mortality rates in Texas and Massachusetts. By looking at each disease independently, we can ensure that the models are fitted to the data distribution without any interaction between the two diseases. We chose to use Poisson regression because it is specifically used to model count data. Poisson regression models the rate of an event's occurrence and provides rate ratios that indicate how changes in uninsured rates influence the likelihood of mortality. After fitting both models, we will compare the results to assess whether the effect of uninsured rates differs between cancer and CHD mortality rates. This comparison will help us understand if the uninsured rate has a stronger or weaker impact on mortality for one disease relative to the other, providing valuable insights.

Upon further review of how our datasets and questions have been organized and cleaned, we found that the planned data evaluation may be unnecessary and infeasible. Since we selected only two states to analyze the impact of disease, we do not have enough data points to create Poisson Regression models. As a result, our data evaluation will have to be done indirectly unless we use all available data points.

2.2 Data Trustworthiness

The US Chronic Disease Indicators dataset (Centers for Disease Control and Prevention., 2024) is sourced from the Centers for Disease Control and Prevention, the national public health agency of the United States and a federal agency under the Department of Health and Human Services (HHS Office of the Secretary and Office of Budget (OB), 2019). Each row consists of a location, a topic, a corresponding question, a data value unit (including raw numbers, cases per 10,000 people, and percentages), type and value, a stratification category, and a corresponding subcategory. The data collected is derived from the American Community Survey, state alcohol sales from the Alcohol Epidemiologic Data System, the American NonSmokers' Rights Foundation database, surveys from the

Behavioral Risk Factor Surveillance System, administrative and claims data from Centers for Medicare & Medicaid Services, the Current Population Survey Food Security Supplement, the National Immunization Survey, the National Survey of Children's Health, census from the National Vital Statistics System, surveys from the Pregnancy Risk Assessment Monitoring System, the U.S. Cancer Statistics Data Visualizations Tool, administrative, claims data and other data from United States Renal Data System, and the Women, Infants, and Children Participant and Program Characteristics Study (Centers for Disease Control and Prevention, 2024).

The Health Insurance Coverage of the Total Population datasets for 2019 (KFF, 2024) are sourced from the Kaiser Family Foundation (KFF), which has been praised for being the “most up-to-date and accurate information on health policy” (Wonkblog Team, 2013). According to the KFF website, “the majority of our health coverage topics are based on analysis of the Census Bureau’s American Community Survey (ACS) by KFF. ACS includes a 1% sample of the US population and allows for precise state-level estimates. The ACS asks respondents about their health insurance coverage at the time of the survey. Respondents may report having more than one type of coverage; however, individuals are sorted into only one category of insurance coverage.” The KFF estimates are derived from the American Community Surveys, with our dataset obtained from 2019 (KFF, 2024).

3. Data Cleaning and View Creation

There are three KFF datasets: one for all adults aged 19–64 and two for males and females aged 19–64. Each dataset has the following values, for the proportion of individuals with the given type of insurance: Location, Employer, Non-Group, Medicaid, Medicare, Military, Uninsured, Total. To clean the files, versions that only include the data of interest with no miscellaneous information are created.

To create the KFF2019_NEW view, since we focus exclusively on uninsured adults, only the Uninsured column is extracted from each dataset. These values are then grouped by location to create the columns All_Uninsured, Female_Uninsured, and Male_Uninsured, representing the proportion of uninsured individuals in each category for each state.

The U.S. Chronic Disease Indicators dataset contains various data types across multiple topics, with 309,215 observations. Given our research questions, and due to the sheer quantity of initial observations, we have created a filtered version of the dataset named USCDI_filter, with Topic as 'Cardiovascular Disease' and 'Cancer', DataValueUnit as 'cases per 100,000' and 'per 100,000', and StratificationCategory1 as 'Sex', 'Age', 'Overall', with the columns down to 10 attributes of interest: YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, and Stratification1.

To create the USCDI view, the column Has2019 is created to determine whether a value is relevant to our analysis. In contrast, Range is created to calculate the average data value (AvgDataValue) across years. This accounts for cases where values are reported over a period greater than one year, under the assumption that DataValue is evenly distributed across the years.

4. Exploratory Data Analysis (EDA)

The USCDI_filter dataset contains 8592 observations and has been narrowed to 10 attributes: YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, and Stratification1. Table 4 presents the selected attributes, along with their descriptions and types. Since the required attributes do not contain missing values (Table 5), imputation is unnecessary.

Impact by Sex

Since CHD mortality in the USCDI_filter dataset or the USCDI view cannot be segregated by both sex and age simultaneously, it is estimated by first obtaining the overall fraction of CHD deaths occurring in females and then applying that fraction to the total number of CHD deaths for individuals aged 0–64, and the results of these calculations are put into a view named USCDI_CHD. We have determined that state- and disease-specific impact data can be obtained directly from the USCDI view.

For the coronary heart disease (CHD) mortality view USCDI_CHD, the USCDI_filter dataset is filtered to include only the relevant cases, with the common unit being USCDI["DataValueUnit"] == "cases per 100,000" and stratified by Sex and Age. Sex is used to estimate the proportion of each gender within each location. This is done by summing the cases per 100,000 people for each gender within a location, regardless of age, and then calculating the proportion of female individuals. Age is used to determine the appropriate age group, with the closest available grouping being the sum of cases per 100,000 people for "Age 0–44" and "Age 45–64". Finally, the proportion of individuals with coronary heart disease is calculated, along with gender-specific proportions, by dividing the values by 100,000.

For the research question “Impact by Sex,” Figures 5 and 6 display bar charts for CHD proportion by location and sex and the uninsurance rate by location and sex, respectively. The CHDPercentage_M values are greater than the corresponding CHDPercentage_F values for both states. However, the difference in morbidity between sexes decreases with age (Lerner & Kannel, 1986). For the uninsurance rate by location and sex, Male_Uninsured values are greater than the corresponding Female_Uninsured values for both states. Figure 7 displays bars representing the ratio of the percentage of uninsured individuals to the percentage of coronary heart disease (CHD) average mortality rates by location and sex, with CHD_Uninsured_Ratio_F values being lower than the corresponding CHD_Uninsured_Ratio_M values for both states. This, combined with the previous two charts, suggests that uninsured females are at a relatively lower risk of CHD average mortality than uninsured males. Since the data from USCDI_CHD and KFF2019_new were already separated by gender during the data cleaning process, there will be minimal changes to how the data is handled.

Impact by State

For the research question "Impact by State," we implemented a Support Vector Regression (SVR) model to analyze average coronary heart disease mortality by state. Since SVR performs poorly with overlapping rows, we addressed this issue by further stratifying the data by age. This stratification ensures that each state has a unique uninsured rate and death rate per age group, reducing redundancy and improving the precision of our analysis. To begin, we separate the features from the target variable. Then, we split the dataset into training and test sets using an 80/20 split.

Additionally, we apply a log transformation to the target variable to reduce skewness and minimize the impact of outliers. This transformation improves the model's fit and enhances interpretability by stabilizing variance and making relationships between variables more linear. A column transformer standardizes the numerical features, while One-Hot Encoding is applied to the location variable to account for state-level variation. These preprocessing steps are integrated into a pipeline to ensure reproducibility and facilitate application across multiple datasets.

Subsequently, hyperparameter optimization is conducted to determine the optimal parameter configuration for the SVR model. Upon identifying the best-performing set of hyperparameters, the model is retrained accordingly. Model performance is evaluated using Root Mean Squared Error (RMSE), which measures predictive accuracy.

Impact by Disease

Finally, regarding the research question “Impact by Disease,” appropriate data filtering and selection were performed. Outlier removal was deemed unnecessary due to the limited number of observations. We selected the necessary attributes to answer this question specifically and renamed some attributes for clarity. Notably, invasive cancer was excluded because it is too broad—it encompasses many different types of "invasive" cancer, making it unsuitable for analyzing the impact of specific cancer types. Outlier detection and removal were not applied to preserve data integrity. A summary table with descriptive statistics for different types of cancer is provided for both states (Table 6). Figures 8 and 9 visualize the cancer types and coronary heart disease comparisons.

5. Results and Discussion

Impact by Sex

The combined table did not provide the number or proportion of females and males in Massachusetts and Texas. Therefore, additional information is needed, and appropriate assumptions must be made. According to the United States Census Bureau, an official website of the U.S. government, the proportion of females is approximately 51% in Massachusetts and 50% in Texas (U.S. Census Bureau., 2025). Since we are using “cases per

100,000" as our data value unit in our analysis, we will assume 51,000 females and 49,000 males in Massachusetts and 50,000 females and 50,000 males in Texas.

In Massachusetts, the total uninsured population is 4.4%. Moreover, 3.1% of females and 5.6% of males are uninsured. Applying adjustments, we estimate that 1,609 females and 2,791 males are uninsured. These figures are calculated as follows:

$$\frac{(0.031\times 51{,}000 + 0.056\times 49{,}000)}{100{,}000} = 0.04325$$

$$Adjustment\ factor = \frac{4{,}400}{4{,}325} = 1.0177$$

$$Adjusted\ female\ uninsured = (0.031 \times 51,000) \times 1.0177 = 1,609$$

$$Adjusted\ male\ uninsured = (0.056 \times 49,000) \times 1.0177 = 2,791$$

The estimated number of deaths in 2019 among individuals aged 0–64 in Massachusetts due to coronary heart disease is 52, of which 16 were females and 36 were males. By dividing the number of deaths by the assumed population sizes (16/51,000 for females and 36/49,000 for males), it is clear that the average mortality rate for males is higher than for females in Massachusetts.

In Texas, the percentage of the population without health insurance is 24.5%. Specifically, 23.2% of females and 25.9% of males are uninsured. After applying adjustments, we estimate that 11,577 females and 12,923 males are uninsured.

$$\frac{(0.232\times 50{,}000 + 0.259\times 50{,}000)}{100{,}000} = 0.2455$$

$$Adjustment\ factor = \frac{24{,}500}{24{,}550} = 0.998$$

$$Adjusted\ female\ uninsured = (0.232 \times 50,000) \times 0.998 = 11,577$$

$$Adjusted\ male\ uninsured = (0.259 \times 50,000) \times 0.998 = 12,923$$

The estimated number of deaths in 2019 among individuals aged 0–64 in Texas due to coronary heart disease is 86, of which 29 were females and 57 were males. By dividing the number of deaths by the assumed population sizes (29/50,000 for females and 57/50,000 for males), it is clear that the average mortality rate for males is higher than for females in Texas.

The average mortality rate of males is higher than that of females in both states. This supports existing research that indicates that CHD incidence and average mortality rates have historically been higher in men than women between the ages of 35 and 84. However, the difference in morbidity between sexes decreases with age (Lerner, Kannel, 1986). Moreover, the uninsured rate for males is higher than for females in both states. This suggests that uninsured females may be at a relatively lower risk of CHD average mortality than uninsured males and that the uninsured rate could influence the average mortality rate in terms of sex.

Impact by Disease

Table 6 shows that the average mortality rates per 100,000 cases for breast cancer, cervical cancer, colorectal cancer, lung cancer, and prostate cancer in Massachusetts are 4.55, 0.28, 2.82, 8.84, and 3.86, respectively. In Texas, the corresponding rates are 4.22, 0.58, 2.74, 6.29, and 2.77. Lung cancer has the highest average mortality rate among all cancer types in both states, followed by breast cancer, prostate cancer, colorectal cancer, and cervical cancer. Massachusetts has higher average mortality rates for breast, colorectal, lung, and prostate cancers, while Texas has a higher average mortality rate for cervical cancer. Overall, the results are similar in both states.

However, notable differences appear in average lung cancer mortality (8.84 in Massachusetts vs. 6.29 in Texas). For coronary heart disease, Massachusetts reports an average mortality rate of 84.0 per 100,000 cases, while Texas reports 88.3. Despite Texas having a significantly higher uninsured rate compared to Massachusetts, the average mortality outcomes are relatively similar, suggesting no clear correlation between average mortality rate and uninsured rate in terms of diseases.

Impact by State

Regression plots were used to visualize relationships, trends, and correlations (Figures 1–3). For the 0–44 and 45–64 age groups, the plots show a clear positive relationship, with the best-fit line indicating that the uninsured rate has predictive power for the average mortality rate. In contrast, for the 65+ age group, the scatter plot shows no clear trend, and the best-fit line has a shallow slope, suggesting that the uninsured rate has limited predictive power for average mortality in this group. Figure 5 presents a visualization of uninsured rates across different states, making comparing and identifying variations and trends easier.

After constructing a Support Vector Regression (SVR) model to analyze coronary heart disease (CHD) average mortality by state, the results show that the best models using states as key support vectors produced test Root Mean Squared Error (RMSE) values of 0.48, 0.41, and 0.19 for the 0–44, 45–64, and 65+ age groups, respectively. The large number of support vectors for *LOCATIONDESC* in the 0–44 and 45–64 age groups suggests that geographic location does not provide strong or consistent differentiation for predicting CHD mortality in these groups. In other words, *LOCATIONDESC* might not be a strong predictor for these age ranges. However, the fewer support vectors for the 65+ group suggest that location may have more predictive value for older adults. This could indicate that geographic factors play a more significant role in CHD mortality for this age group. Rhode Island appears as a support vector for all three age groups, suggesting it has unique characteristics or patterns influencing CHD mortality across ages. This may point to regional factors in Rhode Island that are particularly relevant to the model, regardless of age group.

Overall, the lack of a clear, consistent pattern across all age groups suggests there is no definitive, universal relationship between uninsured rates and CHD mortality across all U.S. states. The influence of *LOCATIONDESC* on CHD mortality appears to vary by age group, with certain states, like Rhode Island, potentially playing a more significant role.

Conclusion and Summary

In Massachusetts, the estimated CHD deaths were 16 for females and 36 for males, yielding higher mortality rates in males when adjusted per 100,000 population. Similarly, the CHD deaths are estimated at 29 for females and 57 for males in Texas. This indicates that males exhibit higher mortality rates. Additionally, the uninsured rate was higher among males in both states. The analysis of different cancer types revealed that lung cancer exhibits the highest mortality rate in both states, followed by breast, prostate, colorectal, and cervical cancers. Massachusetts showed a higher mortality rate overall compared to Texas, specifically, higher mortality rates for breast, colorectal, lung, and prostate cancers. Notably, CHD mortality rates were 84.0 per 100,000 in Massachusetts and 88.3 per 100,000 in Texas. Despite the significantly higher uninsured rate in Texas compared to Massachusetts, the average mortality outcomes for CHD and cancer were similar between the two states. Regression plots indicated a positive relationship between uninsured rates and CHD mortality rates, which supports the fact that the CHD mortality rate is higher in Texas compared to Massachusetts. However, this behaviour diminished in the 65+ age group. SVR models further suggested that geographic location, represented by state-specific support vectors, had limited predictive power for younger age groups but was more influential among older adults. Rhode Island emerged as a notable support vector across all age groups.

This paper comprehensively analyzes the impact of insurance rates from three perspectives: by sex, by disease, and by state, addressing three core research questions. This study presents key findings that provide supporting evidence intended to inform and engage stakeholders in the healthcare and health insurance sectors. However, it is essential to acknowledge that, due to resource limitations, some of the observed outcomes may be attributable to random variation. Furthermore, the study does not account for potential confounding factors that may influence the results, representing a limitation. The analysis primarily addresses the surface-level aspects of the topic; a more in-depth investigation could improve the accuracy and reliability of the findings.

Future research could benefit from expanding the geographic scope, thereby generating a more comprehensive dataset suitable for deeper statistical analysis and more refined age stratifications (e.g., five- or ten-year intervals). Incorporating other confounding factors such as income and educational attainment would enable more precise differentiation between the effects of insurance coverage and broader social determinants of health. Different types of insurance (e.g., private versus public) could also be investigated. Overall, this study lays a foundational basis for future research in health insurance.

6. SQL Script and Schema

6.1 Script

The file knm_datasetup.sql contains the SQL script to load data into the database, which was generated in the Python data-cleaning file. Our SQL scripts are written and yields consistent results as our Python code. Due to the sheer size, the script in knm_datasetup.sql will not be directly posted here.

There are 5 sets of SQL scripts written for the research paper: 2 scripts to create 2 views used for all of the project questions, and 3 scripts for each project question. These can also be found in the README.md of the Github corresponding to this project (Tam, Shiao, Wang, 2025).

KFF2019_NEW View

```
CREATE VIEW KFF2019_NEW AS
SELECT kffa1."Location" AS Location, kffa1."Uninsured" AS All_Uninsured, kfff1."Uninsured" AS Female_Uninsured, kffm1."Uninsured" AS Male_Uninsured
FROM KFF2019_adult kffa1
INNER JOIN KFF2019_female kfff1 ON kffa1."Location" = kfff1."Location"
INNER JOIN KFF2019_male kffm1 ON kffa1."Location" = kffm1."Location"
WHERE kffa1."Location" != 'United States';
```

USCDI View

```
CREATE VIEW USCDI AS
SELECT USCDI_MID."YearStart" AS YearStart,
    USCDI_MID."YearEnd" AS YearEnd,
    USCDI_MID."LocationDesc" AS LocationDesc,
    USCDI_MID."Topic" AS Topic,
    USCDI_MID."Question" AS Question,
    USCDI_MID."DataValueUnit" AS DataValueUnit,
    USCDI_MID."DataValueType" AS DataValueType,
    USCDI_MID."DataValue" AS DataValue,
    USCDI_MID."StratificationCategory1" AS StratificationCategory1,
  USCDI_MID."Stratification1" AS Stratification1,
  USCDI_MID."Has2019" AS Has2019,
  USCDI_MID."Range" AS Range,
  (USCDI_MID."DataValue" / USCDI_MID."Range") AS AvgDataValue
FROM (
    SELECT cdif1."YearStart", cdif1."YearEnd", cdif1."LocationDesc", cdif1."Topic",
cdif1."Question",
      cdif1."DataValueUnit", cdif1."DataValueType", cdif1."DataValue",
      cdif1."StratificationCategory1", cdif1."Stratification1",
      CAST(
        CASE
          WHEN ((cdif1."YearStart" <= 2019) AND (cdif1."YearEnd" >= 2019)) THEN
1
          ELSE 0
        END AS NUMBER(1, 0)
      ) AS "Has2019",
      CAST(
        (cdif1."YearEnd" - cdif1."YearStart" + 1) AS NUMBER(2, 0)
      ) AS "Range"
  FROM USCDI_filter cdif1
  WHERE cdif1."LocationDesc" != 'United States'
) USCDI_MID;
```

Impact by Sex

```
CREATE VIEW USCDI_CHD AS
  WITH CHD_Data AS (
    SELECT
      total."LOCATIONDESC" AS LOCATIONDESC,
               CAST(female.DataValue / (female.DataValue + male.DataValue) AS
DECIMAL(19, 18)) AS Frac_F,
      CAST(total.DataValue AS DECIMAL(24, 18)) AS CHD_DEATHS
    FROM
      (SELECT "LOCATIONDESC", SUM("AVGDATAVALUE") as DataValue
      FROM USCDI
      WHERE "TOPIC" = 'Cardiovascular Disease'
           AND "QUESTION" = 'Coronary heart disease mortality among all people,
underlying cause'
      AND "DATAVALUEUNIT" = 'cases per 100,000'
      AND "STRATIFICATIONCATEGORY1" = 'Age'
      AND "STRATIFICATION1" IN ('Age 0-44', 'Age 45-64')
      AND "DATAVALUETYPE" = 'Crude Rate'
      AND "HAS2019" = 1
      GROUP BY "LOCATIONDESC") total
    JOIN
      (SELECT "LOCATIONDESC", SUM("AVGDATAVALUE") as DataValue
          FROM USCDI
          WHERE "TOPIC" = 'Cardiovascular Disease'
                 AND "QUESTION" = 'Coronary heart disease mortality among all people,
underlying cause'
          AND "DATAVALUEUNIT" = 'cases per 100,000'
          AND "STRATIFICATIONCATEGORY1" = 'Sex'
          AND "STRATIFICATION1" = 'Female'
          AND "DATAVALUETYPE" = 'Age-adjusted Rate'
          AND "HAS2019" = 1
          GROUP BY "LOCATIONDESC") female
       ON total."LOCATIONDESC" = female."LOCATIONDESC"
       JOIN
          (SELECT "LOCATIONDESC", SUM("AVGDATAVALUE") as DataValue
          FROM USCDI
          WHERE "TOPIC" = 'Cardiovascular Disease'
                 AND "QUESTION" = 'Coronary heart disease mortality among all people,
underlying cause'
          AND "DATAVALUEUNIT" = 'cases per 100,000'
          AND "STRATIFICATIONCATEGORY1" = 'Sex'
          AND "STRATIFICATION1" = 'Male'
          AND "DATAVALUETYPE" = 'Age-adjusted Rate'
          AND "HAS2019" = 1
```

```
            GROUP BY "LOCATIONDESC") male
        ON total."LOCATIONDESC" = male."LOCATIONDESC"
    )
    SELECT
        CHD_Data.LOCATIONDESC,
        CHD_Data.FRAC_F,
        CHD_Data.CHD_DEATHS,
        CAST(CHD_Data.CHD_DEATHS * CHD_Data.FRAC_F AS DECIMAL(24, 18)) AS CHD_DEATHS_F,
        CAST(CHD_Data.CHD_DEATHS * (1 - CHD_Data.FRAC_F) AS DECIMAL(24, 18)) AS CHD_DEATHS_M,
        CAST(CHD_Data.CHD_DEATHS / 100000 AS DECIMAL(19, 18)) AS CHDPROP,
        CAST((CHD_Data.CHD_DEATHS * CHD_Data.FRAC_F) / 100000 AS DECIMAL(19, 18)) AS CHDPROP_F,
        CAST((CHD_Data.CHD_DEATHS * (1 - CHD_Data.FRAC_F)) / 100000 AS DECIMAL(19, 18)) AS CHDPROP_M
    FROM CHD_Data


SELECT
    uc.LocationDesc,
    uc.Frac_F,
    uc.CHD_Deaths,
    uc.CHD_Deaths_F,

    uc.CHD_Deaths_M,

    uc.CHDPercentage,

    uc.CHDPercentage_F,

    uc.CHDPercentage_M,

    kff.All_Uninsured,

    kff.Female_Uninsured,

    kff.Male_Uninsured

FROM USCDI_CHD uc

LEFT JOIN KFF2019_new kff

    ON uc.LocationDesc = kff.Location;
```

Impact by State

```
SELECT

    us.LocationDesc,

    us.DataValueUnit AS DeathRateUnit,

    us.DataValueType AS DeathRateType,

    us.AvgDataValue AS AvgDeathRate,

    us.Stratification1,

    kff.All_Uninsured

FROM USCDI us

LEFT JOIN KFF2019_new kff
```

```
    ON us.LocationDesc = kff.Location
WHERE us.Topic = 'Cardiovascular Disease'
      AND us.Question = 'Coronary heart disease mortality among all people, underlying cause'
    AND us.DataValueUnit = 'cases per 100,000'
    AND us.StratificationCategory1 = 'Age'
    AND us.Stratification1 IN ('Age 0-44', 'Age 45-64')
    AND us.DataValueType = 'Crude Rate'
    AND us.Has2019 = 1
    AND us.LocationDesc != 'United States';
```

Impact by Disease

```
SELECT
    us.LocationDesc AS State,
    CASE
        WHEN us.DataValueUnit = 'cases per 100,000' THEN 'per 100,000'
        ELSE us.DataValueUnit
    END AS DeathRateUnit,
    us.DataValueType AS DeathRateType,
    us.AvgDataValue AS AvgDeathRate,
    us.Stratification1,
    us.Question,
    us.DataValue,
    us.Topic AS Disease,
    kff.All_Uninsured
FROM USCDI us
LEFT JOIN KFF2019_new kff
    ON us.LocationDesc = kff.Location
WHERE us.LocationDesc IN ('Texas', 'Massachusetts')
    AND us.Topic IN ('Cardiovascular Disease', 'Cancer')
    AND us.DataValueUnit IN ('cases per 100,000', 'per 100,000')
    AND us.DataValueType = 'Crude Rate'
    AND us.Stratification1 = 'Overall'
    AND us.Has2019 = 1
```

6.2 Schema

KFF2019_ADULT(Location, Employer, Non-Group, Medicaid, Medicare, Military, Uninsured, Total)

KFF2019_FEMALE(Location, Employer, Non-Group, Medicaid, Medicare, Military, Uninsured, Total)

KFF2019_MALE(Location, Employer, Non-Group, Medicaid, Medicare, Military, Uninsured, Total)

USCDI_FILTER(YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, Stratification1)

KFF2019_NEW(Location, All_Uninsured, Female_Uninsured, Male_Uninsured)

USCDI(YearStart, YearEnd, LocationDesc, Topic, Question, DataValueUnit, DataValueType, DataValue, StratificationCategory1, Stratification1, Has2019, Range, AvgDataValue)

USCDI_CHD(LocationDesc, Frac_F, CHD_Deaths, CHD_Deaths_F, CHD_Deaths_M, CHDPercentage, CHDPercentage_F, CHDPercentage_M)

AI Tool Use Declaration

We have used Chegg from Cite This For Me to assist with citations and ChatGPT with Poe for cleaning citations, grammar checking, extracting the libraries used within the process, and helping set up the SQL.

- https://chatgpt.com/share/67e79369-9ddc-8002-9bec-655ee2a3e9f7
- https://chatgpt.com/share/67ead625-40f4-8002-9aae-0bd373633973
- https://chatgpt.com/share/67ed9003-02cc-8002-8866-ebf36a0906e9
- https://chatgpt.com/share/67eef017-aa74-8002-96ea-c50738ee3bb4

References

- Nicholas Tam, Kevin Shiao, and Minghao Wang. 2025. cpsc368_knm_project. (February 2025). Retrieved April 3, 2025 from https://github.com/Nick-2003/cpsc368_knm_project

- Centers for Disease Control and Prevention. 2024. U.S. Chronic Disease Indicators. (March 2024). Retrieved February 9, 2025 from https://healthdata.gov/dataset/U-S-Chronic-Disease-Indicators/dhcp-wb3k/about_data

- KFF. 2024. (October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/adults-19-64/

- HHS Office of the Secretary and Office of Budget (OB). 2019. Centers for Disease Control and Prevention. (November 2019). Retrieved February 9, 2025 from https://web.archive.org/web/20200410150453/https://www.hhs.gov/about/budget/fy-2020-cdc-contingency-staffing-plan/index.html
- Wonkblog Team. 2013. Presenting the third annual WONKY Awards - The Washington Post. (December 2013). Retrieved February 9, 2025 from https://www.washingtonpost.com/news/wonk/wp/2013/12/31/presenting-the-third-annual-wonky-awards/
- KFF. 2024. (October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/health-insurance-coverage-of-women-19-64/
- KFF. 2024. (October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/health-insurance-coverage-of-men-19-64/
- Centers for Disease Control and Prevention. 2024a. Indicator Data Sources. (June 2024). Retrieved February 10, 2025 from https://www.cdc.gov/cdi/about/indicator-data-sources.html
- Junde Li, Qi Ma, Alan HS. Chan, and S.S. Man. 2019. Health monitoring through wearable technologies for older adults: Smart wearables acceptance model. Applied Ergonomics 75 (February 2019), 162–169. DOI: http://dx.doi.org/10.1016/j.apergo.2018.10.006
- Randall R. Bovbjerg and J. Hadley. “Why Health Insurance Is Important,” Urban Institute, 2006. [Online]. Available: https://www.urban.org/sites/default/files/publication/46826/411569-Why-Health-Insurance-Is-Important.PDF. [Accessed: 10-Feb-2025].

- Helen Levy and David Meltzer. 2008. The impact of health insurance on Health. Annual Review of Public Health 29, 1 (April 2008), 399–409. DOI: http://dx.doi.org/10.1146/annurev.publhealth.28.021406.144042
- Lerner, D. J., & Kannel, W. B. (1986). Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. American Heart Journal, 111(2), 383–390. DOI: https://doi.org/10.1016/0002-8703(86)90155-9
- Institute of Medicine (US) Committee on the Consequences of Uninsurance. 1970. Mechanisms and methods: Looking at the impact of health insurance on Health. (January 1970). Retrieved March 31, 2025 from https://www.ncbi.nlm.nih.gov/books/NBK220631/
- Institute of Medicine (US) Committee on the Consequences of Uninsurance. 1970a. Effects of health insurance on Health. (January 1970). Retrieved March 31, 2025 from https://www.ncbi.nlm.nih.gov/books/NBK220636/
- U.S. Census Bureau. 2025. Massachusetts QuickFacts. Retrieved April 2, 2025 from https://www.census.gov/quickfacts/fact/table/MA/SEX255223#SEX255223
- U.S. Census Bureau. 2025. Texas QuickFacts. Retrieved April 2, 2025 from https://www.census.gov/quickfacts/fact/table/TX/PST045224

Tables

Table 1: Selected Attributes with Descriptions and Data Types (KFF datasets (Adults Ages 19-64, Women Ages 19-64, Men Ages 19-64))

Column	Description	Data Type	Property
Location	State within U.S.	VARCHAR2(50)	PRIMARY KEY
Employer	Includes those covered by employer-sponsored coverage either through their own job or as a dependent in the same household.	DECIMAL(19, 18)	N/A
Non-Group	Includes individuals and families that purchased or are covered as a dependent by non-group insurance.	DECIMAL(19, 18)	N/A
Medicaid	Includes those covered by Medicaid, Medical Assistance, Children's Health Insurance Plan (CHIP) or any kind of government-assistance plan for those with low incomes or a disability, as well as those who have both Medicaid and another type of coverage, such as dual eligibles who are also covered by Medicare.	DECIMAL(19, 18)	N/A
Medicare	Includes those covered by Medicare, except dual eligibles who are covered by both Medicaid and Medicare and those covered by Medicare and employer-sponsored insurance who work full-time.	DECIMAL(19, 18)	N/A
Military	Includes those covered under the military or Veterans Administration.	DECIMAL(19, 18)	N/A
Uninsured	Includes those without health insurance and those who have coverage under the Indian Health Service only.	DECIMAL(19, 18)	N/A
Total	Maximum proportion.	DECIMAL(19, 18)	N/A

Table 2: Selected Attributes with Descriptions and Data Types (USDCI_filter)

Column	Description	Data Type	Property
YearStart	Start year of measurements	NUMBER(4, 0)	PRIMARY KEY
YearEnd	End year of measurements	NUMBER(4, 0)	PRIMARY KEY
LocationDesc	State within US	VARCHAR(50)	PRIMARY KEY
Topic	Topic of interest	VARCHAR(30)	PRIMARY KEY
Question	Question of interest, based on Topic	VARCHAR(100)	PRIMARY KEY
DataValueUnit	Unit of data value depending on topic question	VARCHAR(20)	PRIMARY KEY
DataValueType	Type of data value (e.g. Crude value, age-adjusted)	VARCHAR2(20)	PRIMARY KEY
DataValue	Data value, with specific interpretation dependent on its unit, type and topic question	DECIMAL(24, 18)	N/A
StratificationCategory1	Category to stratify data; includes "Age", "Sex", "Race/Ethnicity" and "Overall"	VARCHAR(10)	PRIMARY KEY
Stratification1	Specific group within StratificationCategory1	VARCHAR(10)	PRIMARY KEY

Table 3: Selected Attributes with Descriptions and Data Types (KFF2019_new)

Column	Description	Data Type	Property
Location	State within U.S.	VARCHAR2(50)	PRIMARY KEY
All_Uninsured	Proportion of uninsured	DECIMAL(19, 18)	N/A

	individuals aged between 19 and 64		
Female_Uninsured	Proportion of uninsured female individuals aged between 19 and 64	DECIMAL(19, 18)	N/A
Male_Uninsured	Proportion of uninsured male individuals aged between 19 and 64	DECIMAL(19, 18)	N/A

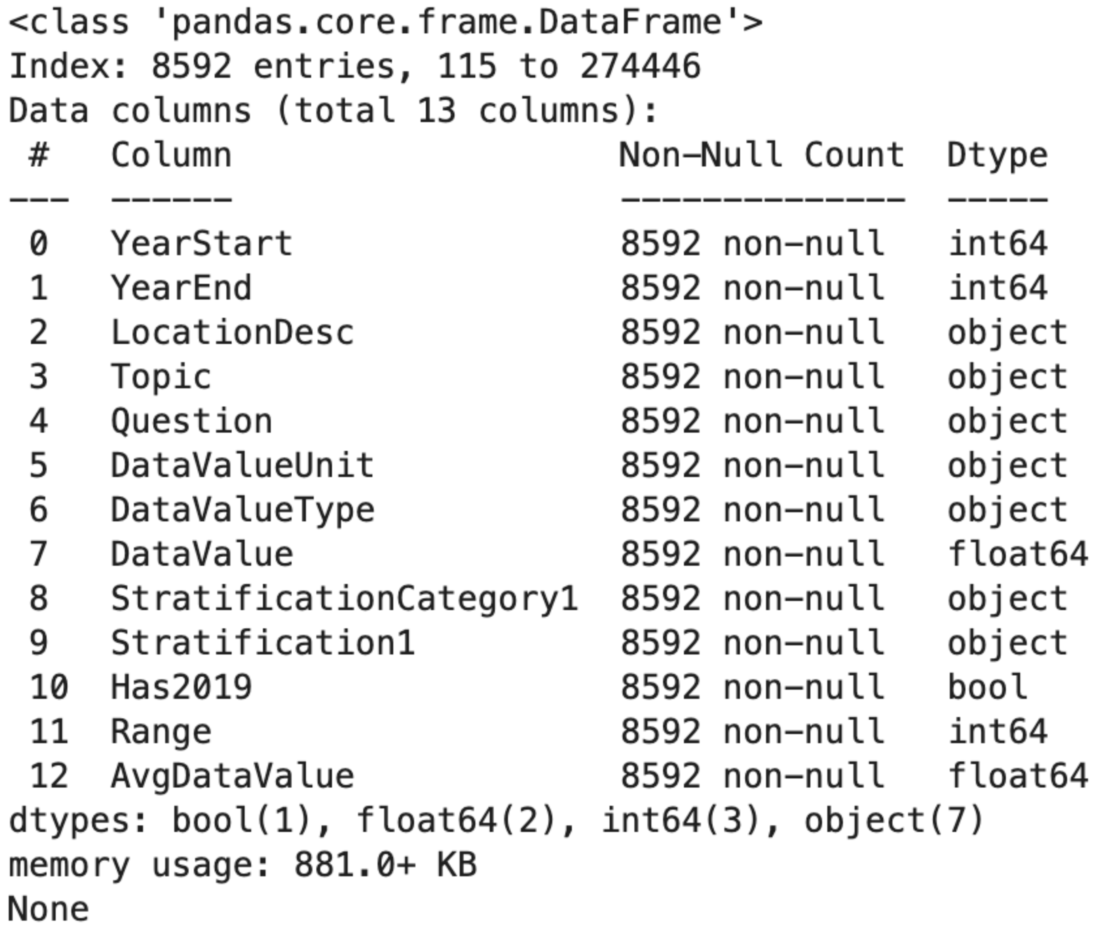
Table 4: Selected Attributes with Descriptions and Data Types (USDCI)

Column	Description	Data Type	Property
YearStart	Start year of measurements	NUMBER(4, 0)	PRIMARY KEY
YearEnd	End year of measurements	NUMBER(4, 0)	PRIMARY KEY
LocationDesc	State within US	VARCHAR(50)	PRIMARY KEY
Topic	Topic of interest	VARCHAR(30)	PRIMARY KEY
Question	Question of interest, based on Topic	VARCHAR(100)	PRIMARY KEY
DataValueUnit	Unit of data value depending on topic question	VARCHAR(20)	PRIMARY KEY
DataValueType	Type of data value (e.g. Crude value, age-adjusted)	VARCHAR2(20)	PRIMARY KEY
DataValue	Data value, with specific interpretation dependent on its unit, type and topic question	DECIMAL(24, 18)	N/A
StratificationCategory1	Category to stratify data; includes “Age”, “Sex”, “Race/Ethnicity”	VARCHAR(10)	PRIMARY KEY

	and “Overall”		
Stratification1	Specific group within StratificationCategory1	VARCHAR(10)	PRIMARY KEY
Has2019	Boolean on whether or not 2019 is in the data	NUMBER(1,0)	NOT NULL
Range	Number of years between YearStart and YearEnd	NUMBER(2,0)	NOT NULL
AvgDataValue	DataValue/Range	DECIMAL(24, 18)	N/A

Table 5: Detection of Missing Values in Required Attributes

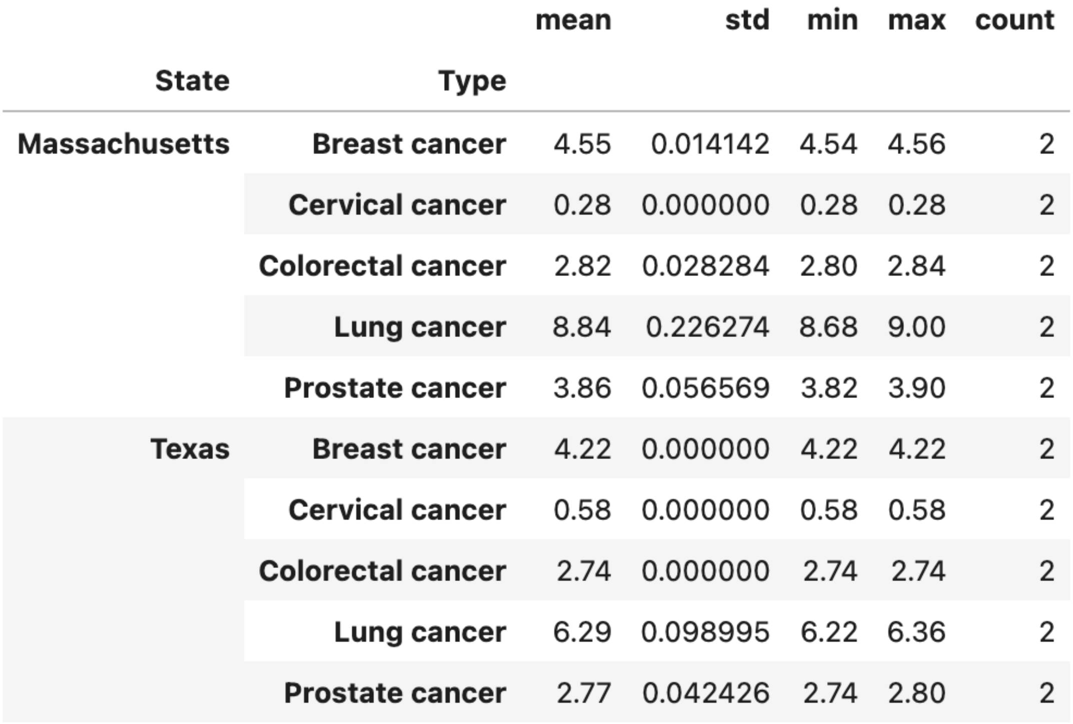
```
display(USCDI.describe())
```



```
<class 'pandas.core.frame.DataFrame'>
Index: 8592 entries, 115 to 274446
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   YearStart                8592 non-null   int64  
 1   YearEnd                  8592 non-null   int64  
 2   LocationDesc             8592 non-null   object 
 3   Topic                    8592 non-null   object 
 4   Question                 8592 non-null   object 
 5   DataValueUnit            8592 non-null   object 
 6   DataValueType            8592 non-null   object 
 7   DataValue                8592 non-null   float64
 8   StratificationCategory1  8592 non-null   object 
 9   Stratification1          8592 non-null   object 
 10  Has2019                  8592 non-null   bool   
 11  Range                    8592 non-null   int64  
 12  AvgDataValue             8592 non-null   float64
dtypes: bool(1), float64(2), int64(3), object(7)
memory usage: 881.0+ KB
None
```

Table 6: Summary Table for Different Types of Cancer with Descriptive Statistics

```
CANCER.groupby(["State","Type"])["AvgDeathRate"].agg(["mean","std","min","max","count"]))
```



		mean	std	min	max	count
State	**Type**					
Massachusetts	**Breast cancer**	4.55	0.014142	4.54	4.56	2
	Cervical cancer	0.28	0.000000	0.28	0.28	2
	Colorectal cancer	2.82	0.028284	2.80	2.84	2
	Lung cancer	8.84	0.226274	8.68	9.00	2
	Prostate cancer	3.86	0.056569	3.82	3.90	2
Texas	**Breast cancer**	4.22	0.000000	4.22	4.22	2
	Cervical cancer	0.58	0.000000	0.58	0.58	2
	Colorectal cancer	2.74	0.000000	2.74	2.74	2
	Lung cancer	6.29	0.098995	6.22	6.36	2
	Prostate cancer	2.77	0.042426	2.74	2.80	2

Figures

Figure 1: Average Death Rate by Uninsured Rate (Age: 0 - 44)

```
sns.regplot(data=state_df_0_44, x='All_Uninsured', y='AvgDeathRate', scatter=True)
plt.title('Uninsured Rate vs Death Rate Age: 0-44')
plt.show()
```

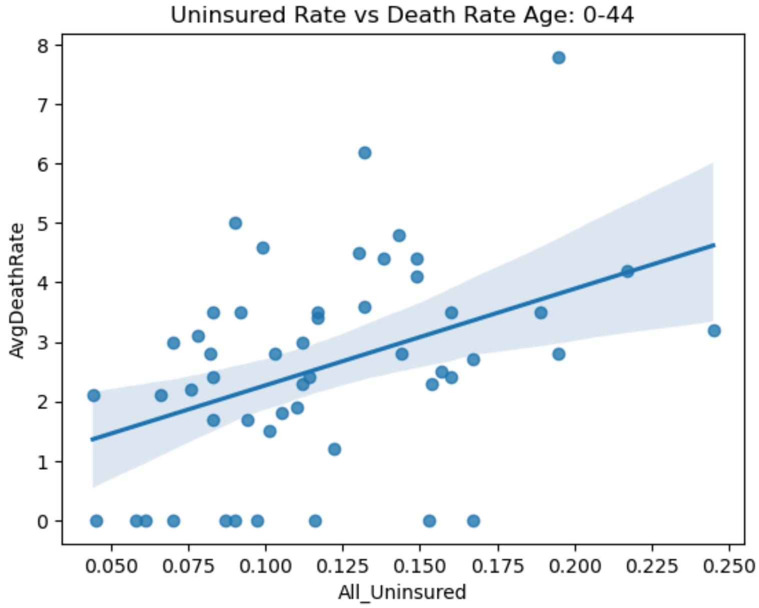


Figure 2: Average Death Rate by Uninsured Rate (Age: 44 - 64)

```
sns.regplot(data=state_df_45_64, x='All_Uninsured', y='AvgDeathRate', scatter=True)
plt.title('Uninsured Rate vs Death Rate Age: 44-64')
plt.show()
```

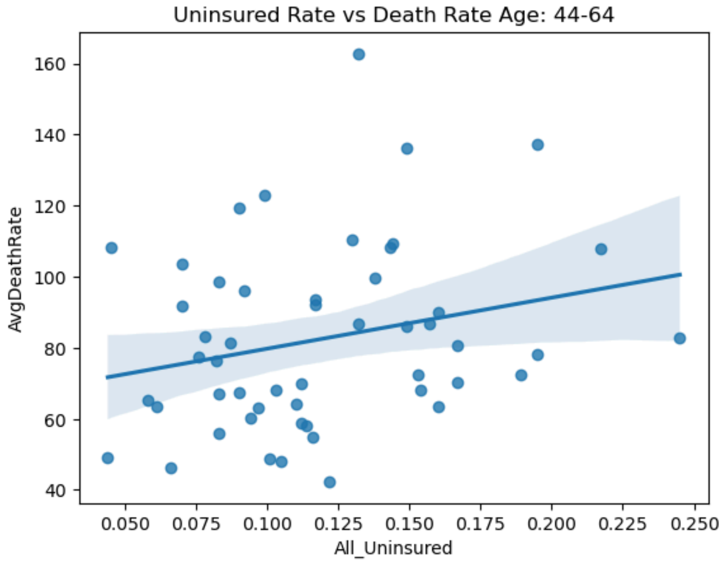


Figure 3: Average Death Rate by Uninsured Rate (Age: 65+)

```
sns.regplot(data=state_df_65, x='All_Uninsured', y='AvgDeathRate', scatter=True)
plt.title('Uninsured Rate vs Death Rate Age: 65+')
plt.show()
```

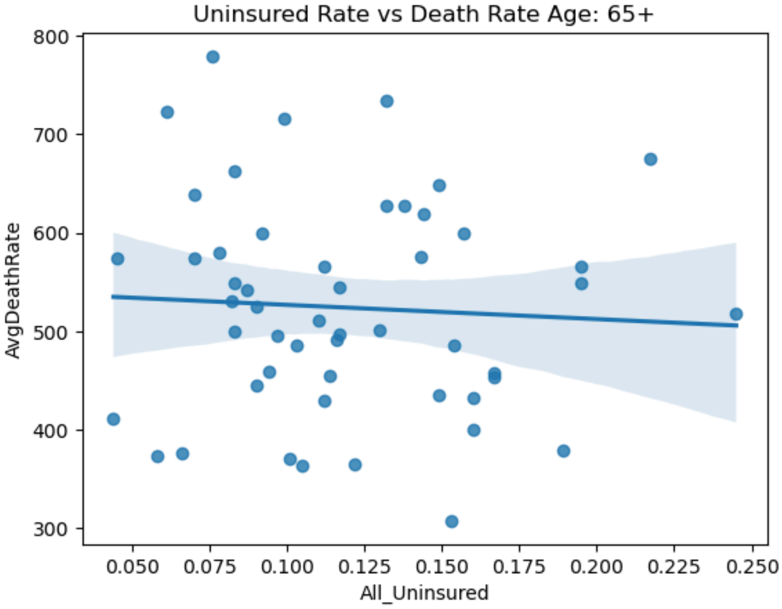


Figure 4: Uninsured Rate by State with Average Uninsured Rate

```
average_uninsured_rate = state_df_65['All_Uninsured'].mean()
plt.figure(figsize=(12, 6))
sns.stripplot(data=state_df_65, x="LocationDesc", y="All_Uninsured", jitter=True, palette="Set2", alpha=0.7)
plt.axhline(y=average_uninsured_rate, color='blue', linestyle='--', label=f'Avg Uninsured Rate: {average_uninsured_rate:.2f}')
plt.title("Uninsured Rate by State with Average Uninsured Rate")
plt.xlabel("State")
plt.ylabel("Uninsured Rate")
plt.xticks(rotation=90)
plt.legend()
plt.show()
```

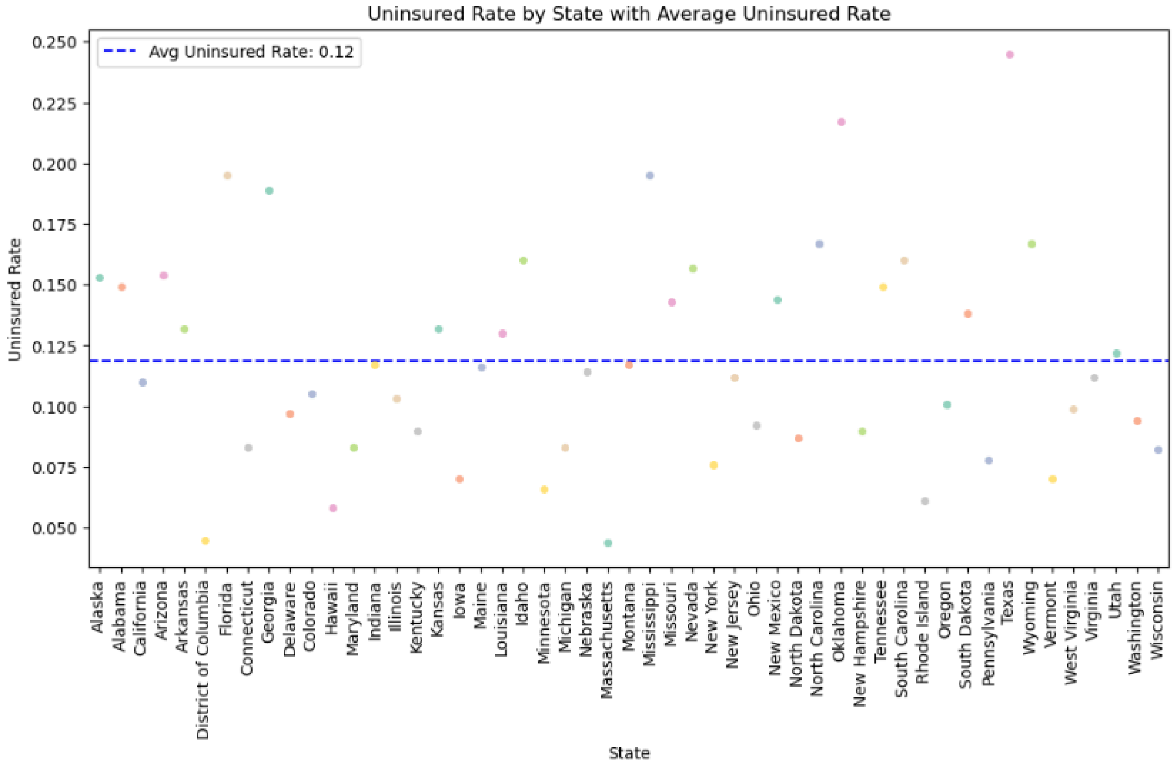


Figure 5: Bar Chart: CHDPercentage by Location and Sex

```
alt.Chart(total_data_focus).mark_bar().encode(
    alt.Y("LocationDesc:N"),
    alt.X(alt.repeat('row'),
          type='quantitative',
          scale=alt.Scale(domain=[0,max(total_data_focus['CHDPercentage_F'].max(),
total_data_focus['CHDPercentage_M'].max())]))
).repeat(
  row=['CHDPercentage_F', 'CHDPercentage_M',]
).properties(
    title="CHD Percentage for Females and Males by Location"
)
```

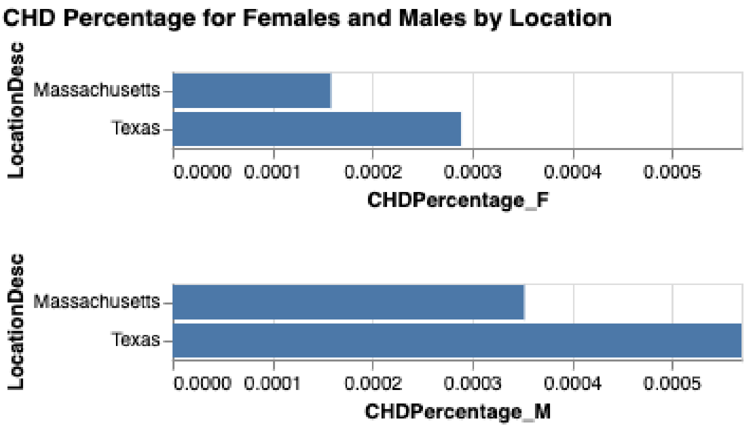


Figure 6: Bar Chart: Uninsurance rate by Location and Sex

```
alt.Chart(total_data_focus).mark_bar().encode(
    alt.Y("LocationDesc:N"),
    alt.X(alt.repeat('row'),
          type='quantitative',
          scale=alt.Scale(domain=[0,max(total_data_focus['Female_Uninsured'].max(),
total_data_focus['Male_Uninsured'].max())]))
).repeat(
  row=['Female_Uninsured', 'Male_Uninsured',]
).properties(
    title="Percentage of Uninsured Individuals for Females and Males by Location"
)
```

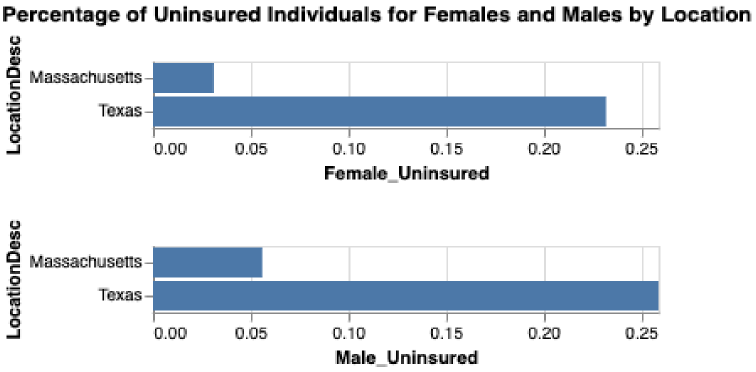


Figure 7: Bar Chart: Ratio of the Percentage of Uninsured Individuals Over the Percentage of Coronary Heart Disease (CHD) Mortality Rates by Location and Sex

```
total_data_focus["CHD_Uninsured_Ratio_F"] = total_data_focus["CHDPercentage_F"] /
total_data_focus["Female_Uninsured"]
total_data_focus["CHD_Uninsured_Ratio_M"] = total_data_focus["CHDPercentage_M"] /
total_data_focus["Male_Uninsured"]
alt.Chart(total_data_focus).mark_bar().encode(
    alt.Y("LocationDesc:N"),
    alt.X(alt.repeat('row'),
          type='quantitative',
          scale=alt.Scale(domain=[0,max(total_data_focus['CHD_Uninsured_Ratio_F'].max(),
total_data_focus['CHD_Uninsured_Ratio_M'].max())]))
).repeat(
  row=['CHD_Uninsured_Ratio_F', 'CHD_Uninsured_Ratio_M',]
).properties(
    title="Ratio of CHD Mortality Percentage over Uninsured Percentage for Females and Males by
Location"
)
```

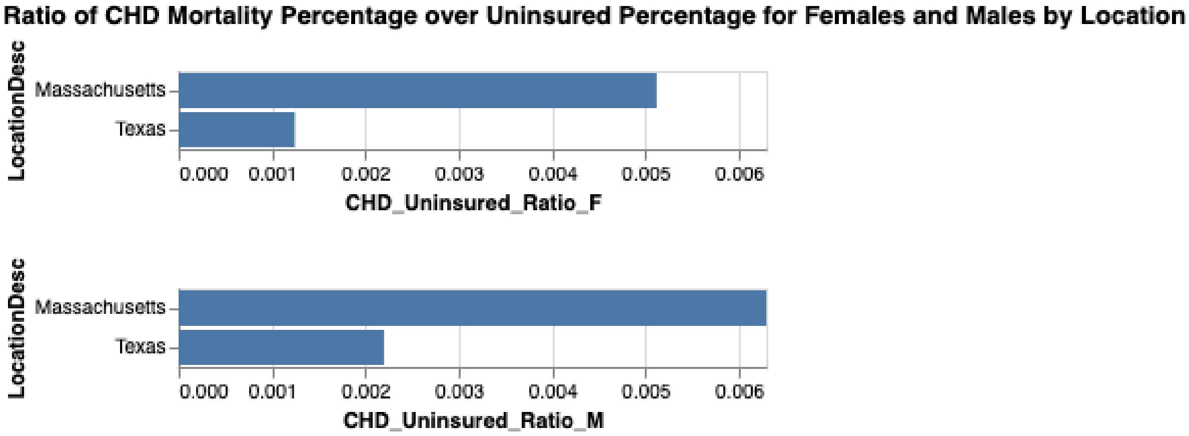


Figure 8: Cancer Types Comparison

```
plt.figure(figsize=(8, 5))
sns.barplot(data= CHD, x='Type', y='AvgDeathRate', hue='State')
plt.xlabel('')
plt.ylabel('Average Death Rate')
plt.title('Coronary Heart Disease Comparison')
plt.legend(title='State')
plt.show()
```

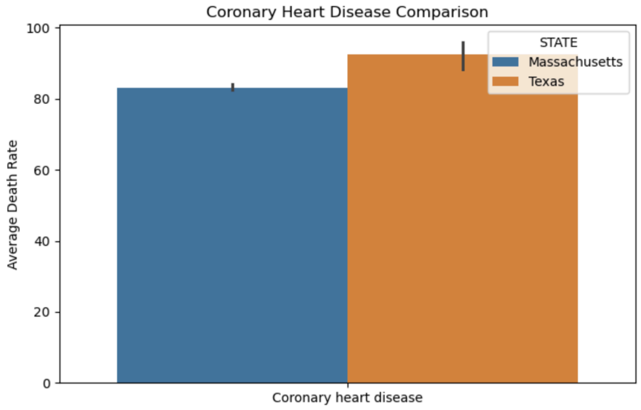
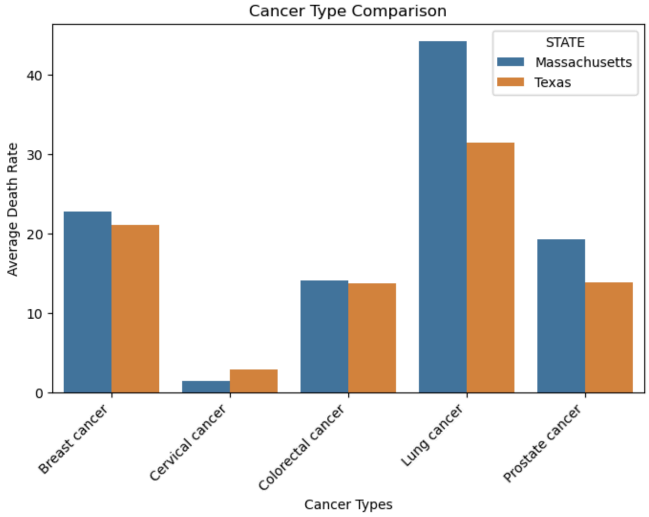


Figure 9: Cancer Types Comparison

```
eda_summary_cancer = CANCER.groupby(['State',
'Type'])['AvgDeathRate'].mean().reset_index()
plt.figure(figsize=(8, 5))
sns.barplot(data=eda_summary_cancer, x='Type', y='AvgDeathRate', hue='State')
plt.xticks(rotation=45, ha='right')
plt.xlabel('Cancer Type')
plt.ylabel('Average Death Rate')
plt.title('Cancer Type Comparison')
plt.legend(title='State')
plt.show()
```


Cancer Type Comparison
Average Death Rate
STATE
Massachusetts
Texas
Breast cancer
Cervical cancer
Colorectal cancer
Lung cancer
Prostate cancer
Cancer Types