CPSC 368 Research Project Midway Checkpoint

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

1. Summary

As Levy and Meltzer suggest, determining whether health insurance plays a significant role in influencing health will likely require substantial investment in social experiments (Levy, Meltzer, 2008). Although this research 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 through the datasets ‘U.S. Chronic Disease Indicators’ and ‘Health Insurance Coverage of Adults Ages 19-64’ from HealthData.gov and KFF, respectively.

2. Research Questions

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 between Texas and Massachusetts; (2) by state, examining coronary heart disease mortality across different states; and (3) by disease, comparing coronary heart disease mortality and various cancer mortalities between Texas and Massachusetts. After consulting with the TA, we have deemed our research questions appropriate for our analysis and have not changed them from the initial proposal. However, 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, except for the analysis by state. Since we selected only two states to analyze the impact of sex and disease, we do not have enough data points to conduct Multiple Linear Regression or create Poisson Regression models. As a result, our data evaluation will have to be done indirectly unless we use all available data points.

3. Data Cleaning

There are three KFF datasets: one for all adults aged 19–64 and two for males and females aged 19–64. Each dataset has a corresponding Group column applied before they are joined on Location. Since our focus is 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 county.

The U.S. Chronic Disease Indicators (USCDI) dataset contains various data types across multiple topics, to a total of 309,215 observations in total. Given our research questions, and due to the sheer quantity of initial observations, we need to create 2 new datasets: a USCDI dataset filtered with Topic as 'Cardiovascular Disease' and 'Cancer', DataValueUnit as 'cases per 100,000' and 'per 100,000', and StratificationCategory1 as 'Sex', 'Age', 'Overall', as well as a dataset for coronary heart disease (CHD) mortality by sex for individuals aged 0–64. The former dataset contains all the attributes that we require for our project questions. As for the latter, since CHD mortality 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. We have determined that state- and disease-specific impact data can be obtained directly from the filtered USCDI dataset. For the overall dataset, 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.

For the coronary heart disease (CHD) mortality dataset, the filtered USCDI 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.

4. Exploratory Data Analysis (EDA)

The filtered USCDI dataset contains 8592 observations and has been narrowed down to 13 attributes. The selected attributes, along with their descriptions and types, are presented in Table 1. Since the required attributes do not contain missing values (Table 2), imputation is unnecessary.

For the research question "Impact by State," we implemented a Support Vector Regression (SVR) model to analyze 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 illustrate this relationship effectively, we used regression plots, as they provide a clear visual representation of trends and correlations (Figure 1 - 3). For the 0–44 and 45–64 age groups, the regression plots show a clear positive relationship, with the best-fit line indicating that the uninsured rate has predictive power for the death rate. In contrast, for the 65+ age group, the scatter plot lacks a clear trend, and the best-fit line has a shallow slope, suggesting that the uninsured rate has limited predictive power for the death rate. However, we will explore incorporating state-level factors and evaluate how a more complex model, such as SVR, performs. The visualization of uninsured rate across different states is presented in Figure 5. This visualization allows us to easily compare each state's uninsured rate, highlighting variations and trends between states.

For the research question "Impact by Sex," Figures 5 and 6 display bar charts for CHD percentage by location and sex, and for the uninsurance rate by location and sex, respectively. The CHDPercentage_M values are greater than the corresponding CHDPercentage_F values for both states. This supports existing research indicating that CHD incidence and mortality rates have historically been higher in men than in women between the ages of 35 and 84, though 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) 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 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.

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. To preserve data integrity, outlier detection and removal were not applied. A summary table with descriptive statistics for different types of cancer is provided for both states (Table 3). According to the table, lung cancer has the highest average death rate among all cancer types in both states, followed by breast cancer, prostate cancer, colorectal cancer, and cervical cancer. The results are similar in both states, except for lung cancer (8.84 in Massachusetts vs. 6.29 in Texas) and prostate cancer (3.86 in Massachusetts vs. 2.77 in Texas), where Massachusetts has a higher average death rate than Texas. For coronary heart disease, Massachusetts reports an average death rate of 84.0, while Texas reports 88.3. Figures 8 and 9 visualize the comparisons for cancer types and coronary heart disease, respectively.

5. SQL Script and Schema

Script

The file knm_datasetup.sql contains the SQL script to load data into the database. Due to the sheer size, the script itself will nto be posted here.

Schema

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)

KFF2019_new(Location, All_Uninsured, Female_Uninsured, Male_Uninsured)

AI Tool Use Declaration

We have used Chegg from Cite This For Me to assist with citations, ChatGPT and Poe for grammar checking and data cleaning.

- https://poe.com/s/zbH24rcNHMAwHjFo1t4S
- https://poe.com/s/aJDH3smuLfVLzbgg8B6y
- https://poe.com/s/k1DzuYValJfK4fHh0FkK
- https://chatgpt.com/share/67cf3582-cf08-8002-aa48-1ee2ae818d2b

References

- 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/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

- 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/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- KFF. 2024.(October 2024). Retrieved February 10, 2025 from https://www.kff.org/other/state-indicator/health-insurance-coverage-of-men-19-64/?currentTimeframe=3&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- 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. https://doi.org/10.1016/0002-8703(86)90155-9

Tables

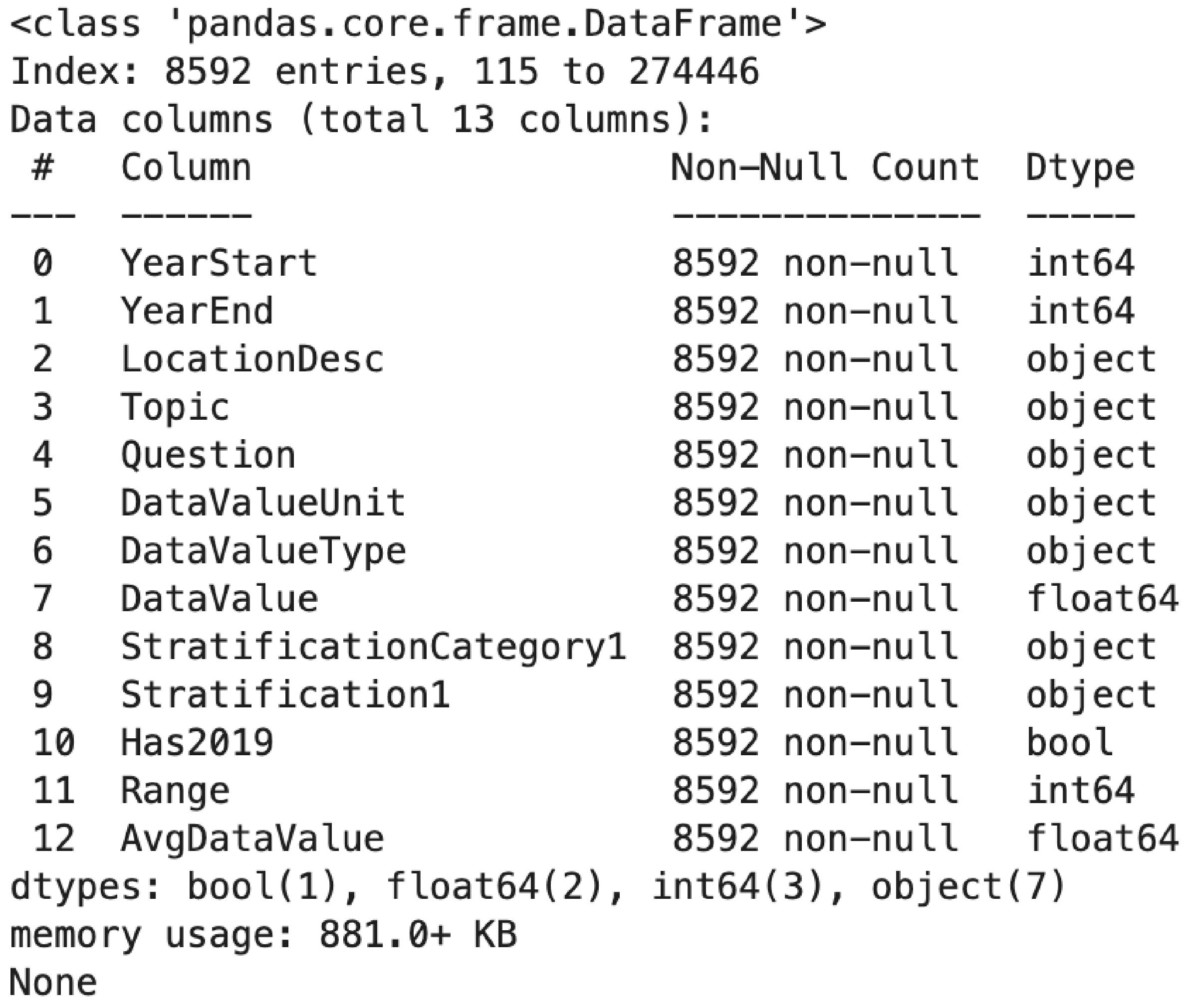
Table 1: Selected Attributes with Descriptions and Data Types

Column	Description	Data Type
YearStart	Start year of measurments	NUMERIC
YearEnd	End year of measurements	NUMERIC
LocationDesc	State within US	VARCHAR
Topic	Topic of interest	VARCHAR
Question	Question of interest, based on Topic	VARCHAR
DataValueUnit	Unit of data value depending on topic question	VARCHAR
DataValueType	Type of data value (e.g. Crude value, age-adjusted)	VARCHAR
DataValue	Data value, with specific interpretation dependent on its unit, type and topic question	NUMERIC
StratificationCategory1	Category to stratify data; includes “Age”, “Sex”,	VARCHAR

	“Race/Ethnicity” and “Overall”	
Stratification1	Specific group within StratificationCategory1	VARCHAR
Has2019	Boolean on whether or not 2019 is in the data	BOOLEAN
Range	Number of years between YearStart and YearEnd	NUMERIC
AvgDataValue	DataValue/Range	NUMERIC

Table 2: Detection of Missing Values in Required Attributes

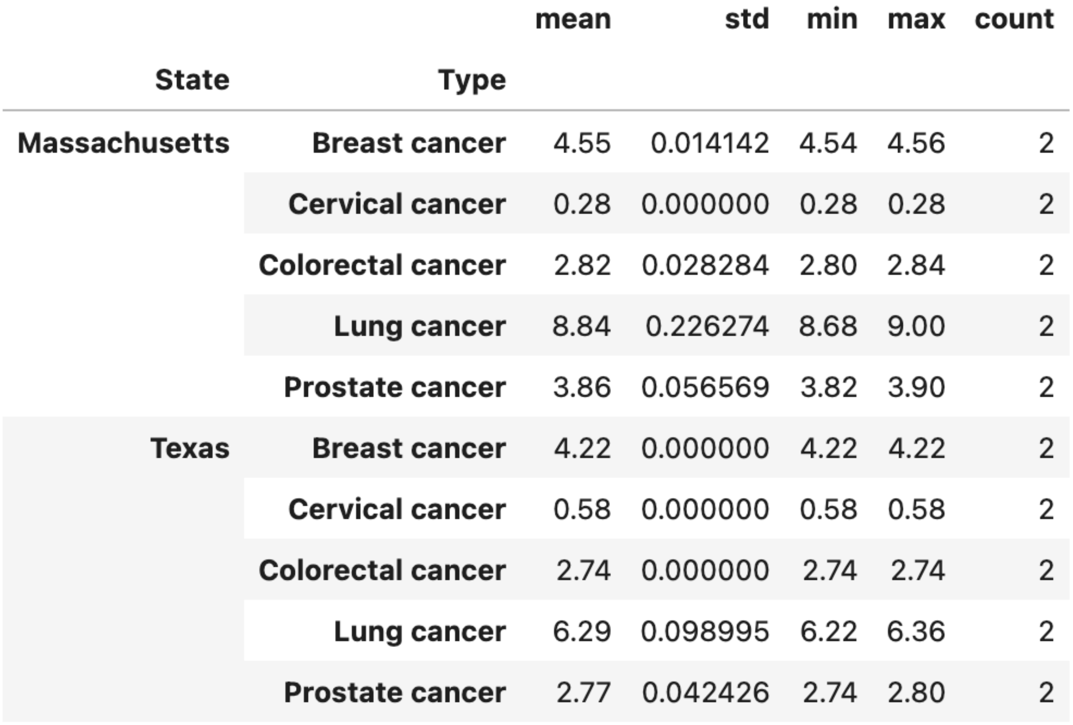
```
display(USCDI.describe())
```



```
<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 3: 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()
```

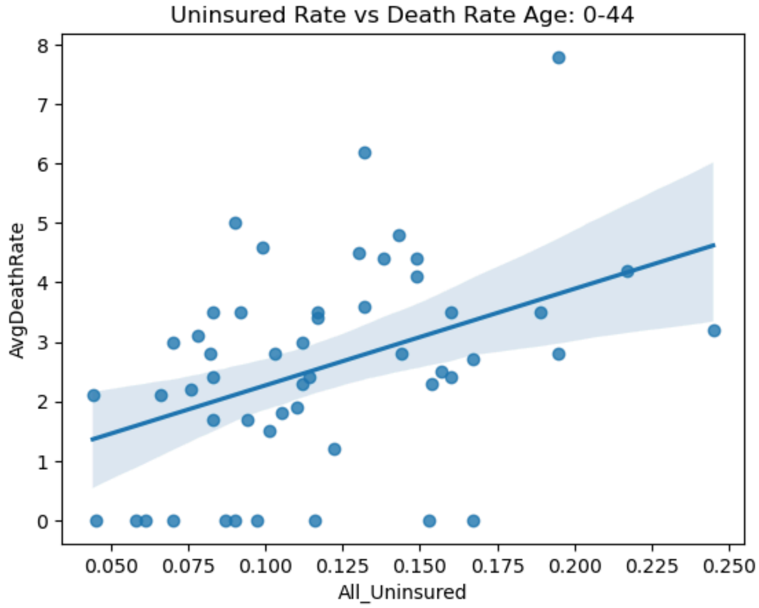


Figure 2: Average Death Rate by Uninsured Rate (Age: 44 - 64)

```
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()
```

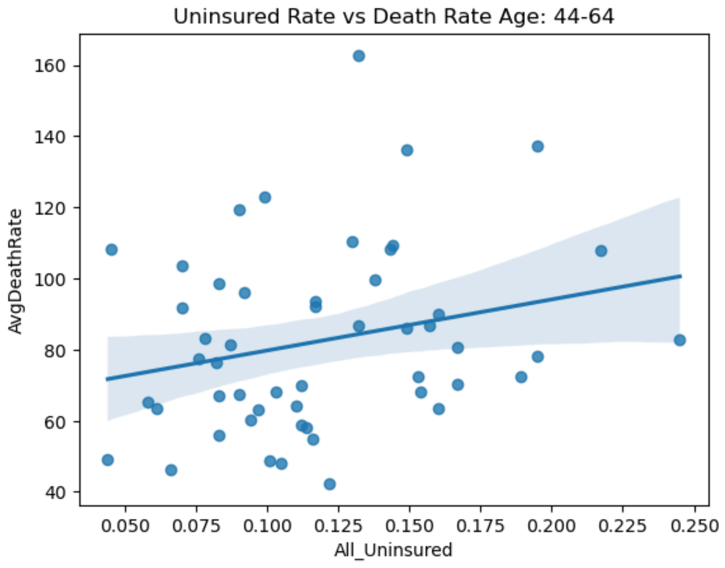


Figure 3: Average Death Rate by Uninsured Rate (Age: 65+)

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

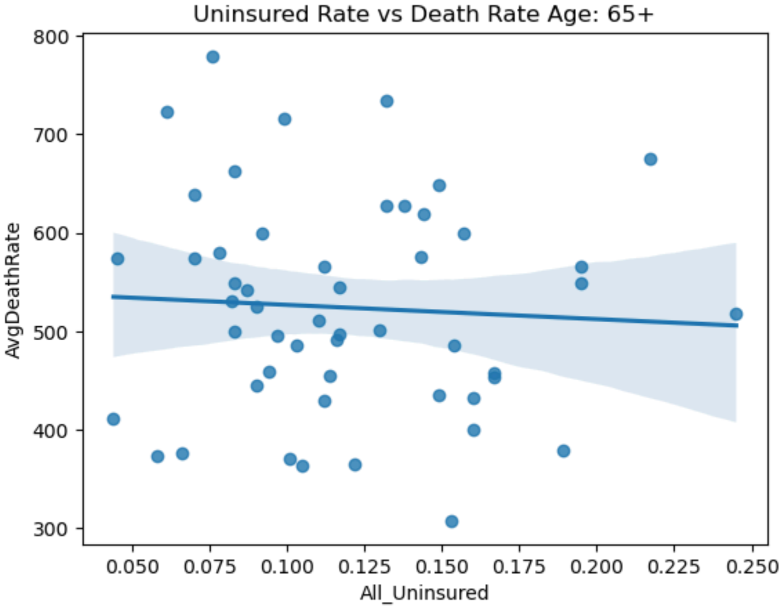


Figure 4: Uninsured Rate by State with Average Uninsured Rate

```
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()
```

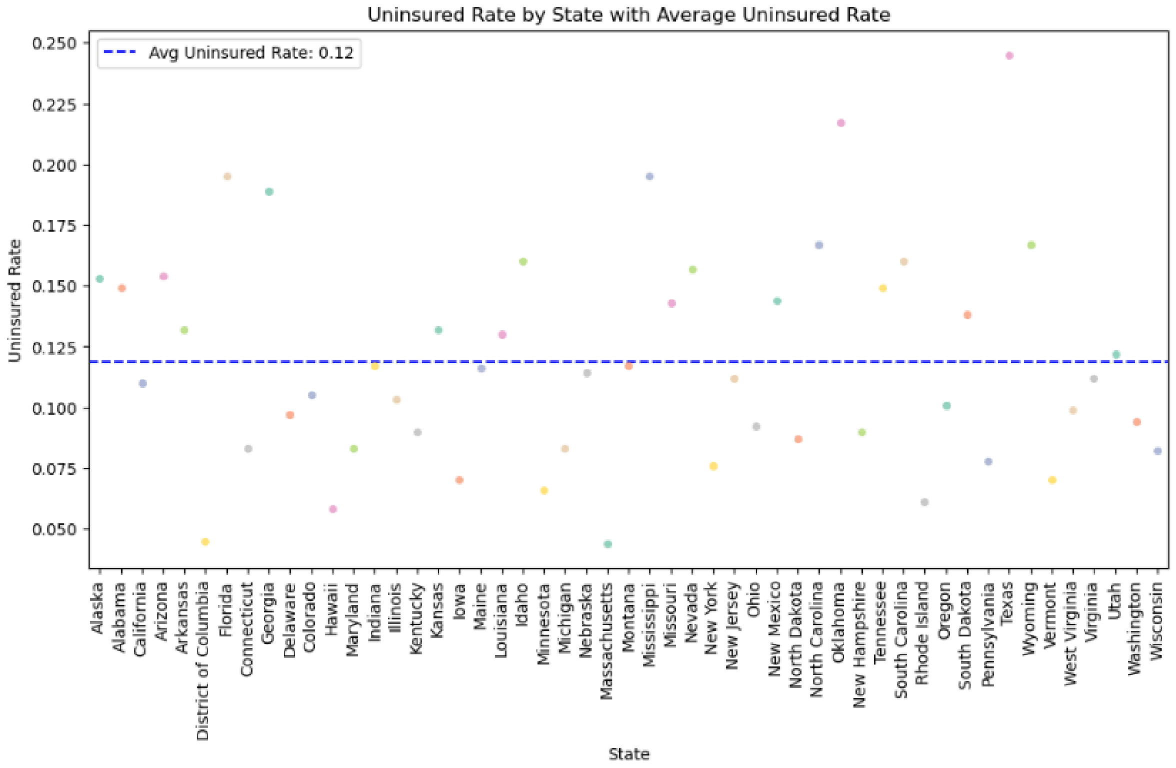


Figure 5: Bar Chart: CHDPercentage by Location and Sex

```
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"
)
```

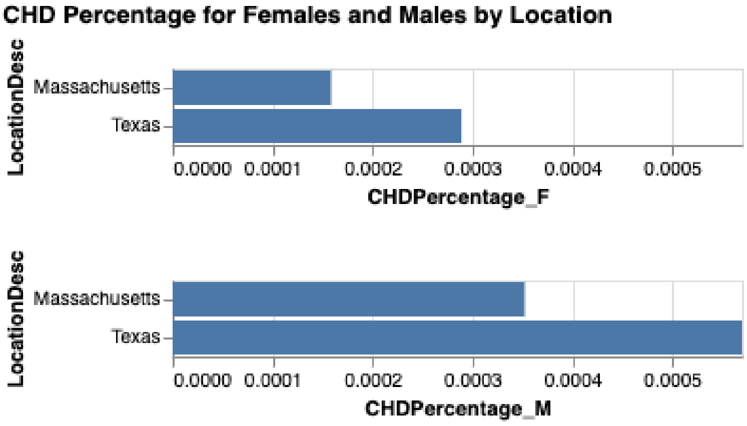


Figure 6: Bar Chart: Uninsurance rate by Location and Sex

```
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"
)
```

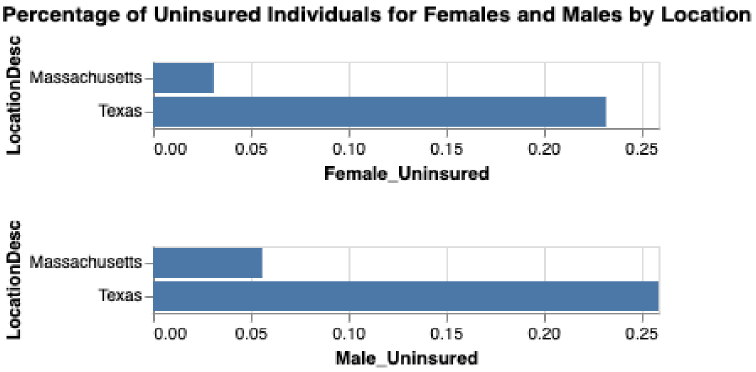


Figure 7: Bar Chart: Ratio of the Percentage of Uninsured Individuals Over the Percentage of Coronary Heart Disease (CHD) Mortality Rates by Location and Sex

```
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"
)
```

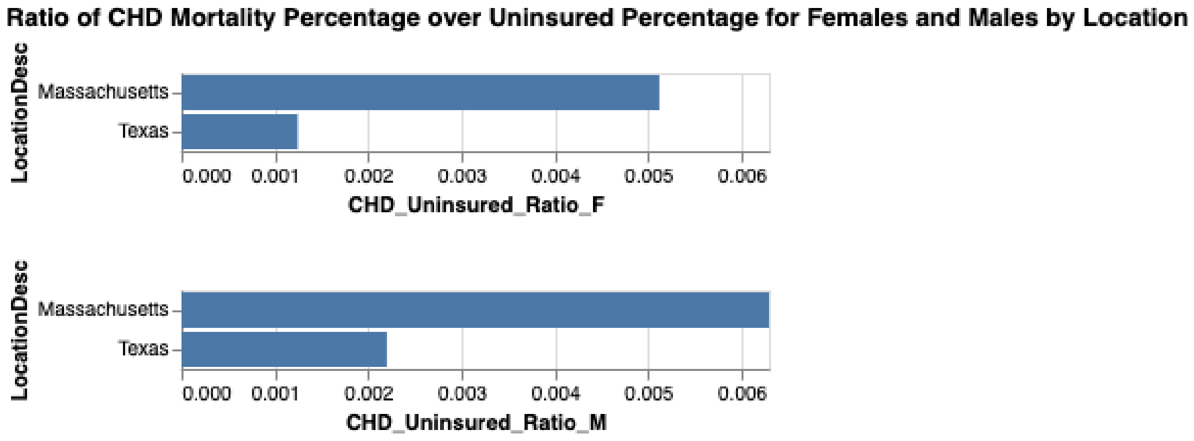

Ratio of CHD Mortality Percentage over Uninsured Percentage for Females and Males by Location

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()
```

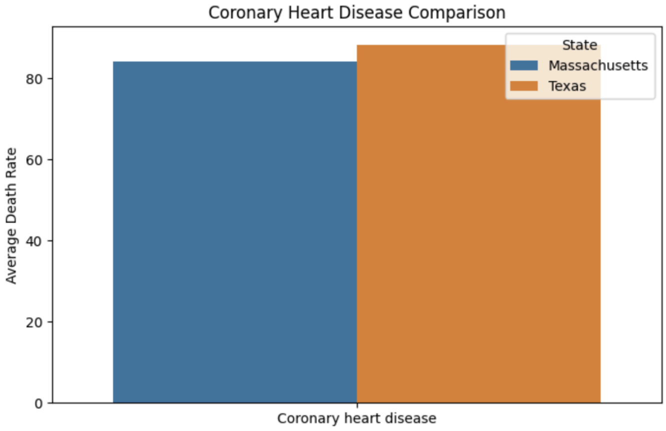
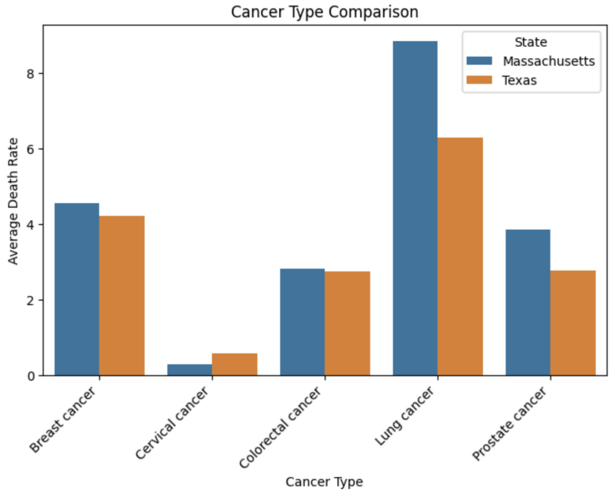


Figure 9: Cancer Types Comparison

```
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
State
Massachusetts
Texas
Average Death Rate
Breast cancer
Cervical cancer
Colorectal cancer
Lung cancer
Prostate cancer
Cancer Type