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 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; (2) by state, examining coronary heart disease mortality across different states; and (3) by disease, comparing coronary heart disease mortality with various cancer mortalities.

Given the lack of feedback from the TA, we presume that the research questions will be optimal for our analysis and thus have not been changed from the initial proposal.

3. Data Cleaning

There are 3 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 to them before they are joined on Location. Since our focus is on uninsured adults exclusively, only the Uninsured column of values is acquired for each individual dataset, which are then grouped by location to create the columns All_Uninsured, Female_Uninsured, and Male_Uninsured, corresponding to the proportion of uninsured individuals in each category for each country.

The U.S. Chronic Disease Indicators dataset contains many types of data for a variety of topics, and given our topic questions, we will create 3 datasets: one for coronary heart disease mortality by gender, one for coronary heart disease by state, and one for the average of various cancer mortalities. The column Has2019 is created to determine if the value is relevant to our questions. In contrast, Range is created to help provide the average data value AvgDataValue across years, given that some values are obtained for a range greater than 1 year.

For the coronary heart disease mortality dataset, the U.S. Chronic Disease Indicators dataset is filtered for the corresponding cases, with the common unit being "USCDI["DataValueUnit"] == cases 'per 100,000'" and with the stratification categories of Sex and Age. Sex is used to estimate the proportion of each gender within each location. This is achieved by obtaining the sum of cases per 100,000 people for each location and gender, regardless of age, followed by calculating the proportion of female individuals present. Age is used to get the appropriate age group, with the closest achievable groups being the sum of cases per 100,000 people between "Age 0-44" and "Age 45-64". Finally, the proportion of individuals that had coronary heart disease is calculated, along with the corresponding

proportions for each gender, by dividing their values by 100000. The column “AvgDataValue” is renamed “CHD_Deaths” to make future interpretation easier for users.

For the cancer dataset, the U.S. Chronic Disease Indicators dataset is filtered for the corresponding cases with data including 2019, with the common unit being “USCDI["DataValueUnit"] == 'per 100,000'” and with the stratification category Sex, as the category Age is not provided. The “Female” and “Male” columns are renamed “Cancer_Deaths_F” and “Cancer_Deaths_M” respectively, to make interpretation easier for future users. The proportions of individuals that acquired some form of cancer are then calculated by dividing the corresponding values by 100000.

4. Exploratory Data Analysis (EDA)

After reading the data file, we see that it contains 309,215 observations and 13 attributes.

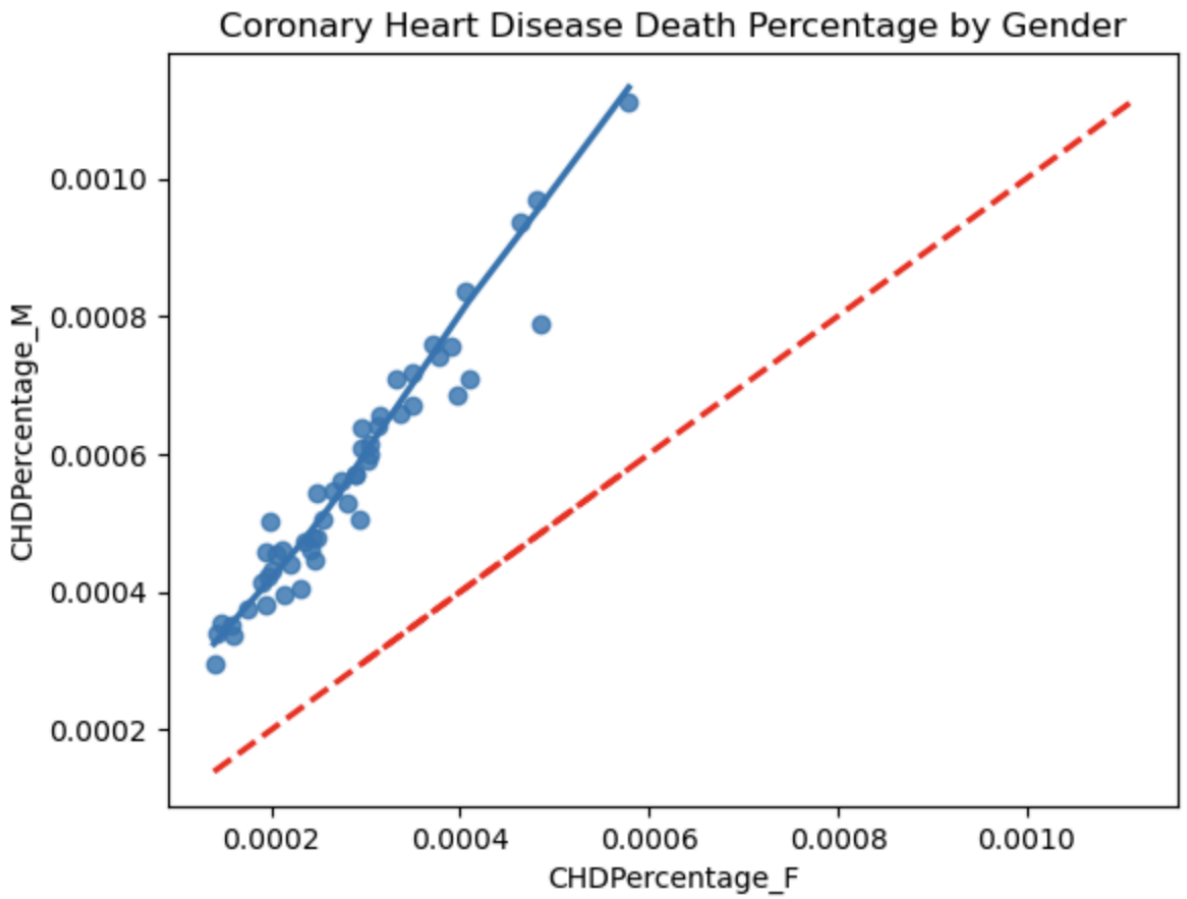
Code: print(state_df.isnull().sum())

The attributes needed from the final dataset does not contain missing values, therefore imputation is unnecessary.

```
Stratification1    0
Type               0
State              0
Disease            0
DeathRateUnit      0
DeathRateType      0
AvgDeathRate       0
All_Uninsured      0
dtype: int64
```

The visualisation below plots the percentage of female coronary heart disease mortalities against the percentage of female coronary heart disease mortalities, grouped together by country, to examine the difference in coronary heart disease mortalities by sex. The visualisation shows that for all countries, the CHDPercentage_M value is consistently greater than the corresponding CHDPercentage_F value. This supports existing research that

indicates that CHD incidence and mortality rates have historically been higher in men than women between the ages 35 and 84, though the difference in morbidity between sexes decreases with age (Lerner, Kannel, 1986).



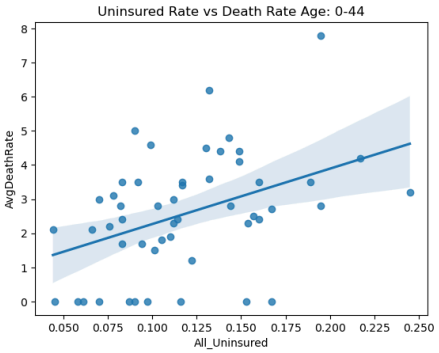
Given the results above, we will later need to use multiple linear regression to explore the relationship between the rate of uninsured individuals and coronary heart disease mortalities and how sex influences said relationship. Since the data from USCDI_CHD and KFF2019_new have already been separated by gender through the cleaning process, there is little change to how the data will be handled.

To analyze coronary heart disease mortality by state, we implemented a Support Vector Regression (SVR) model. Since SVR performs poorly with overlapping rows, we addressed this issue by further stratifying the data by age. This ensures that each state has a unique uninsured rate and death rate per age group, reducing redundancy and enhancing the

precision of our analysis. Below is the code to visualize the average death rate by uninsured rate. To illustrate this relationship effectively, I've chosen to use regression plots, as they provide a clear visual representation of trends and correlations.

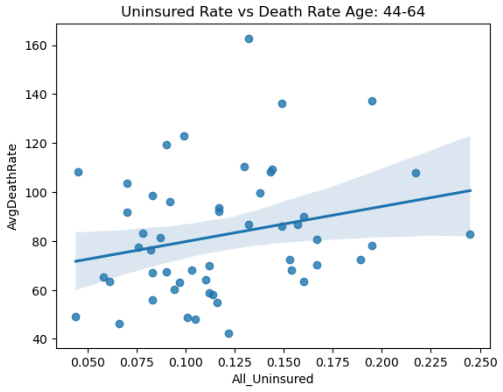
Code

```
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()
```



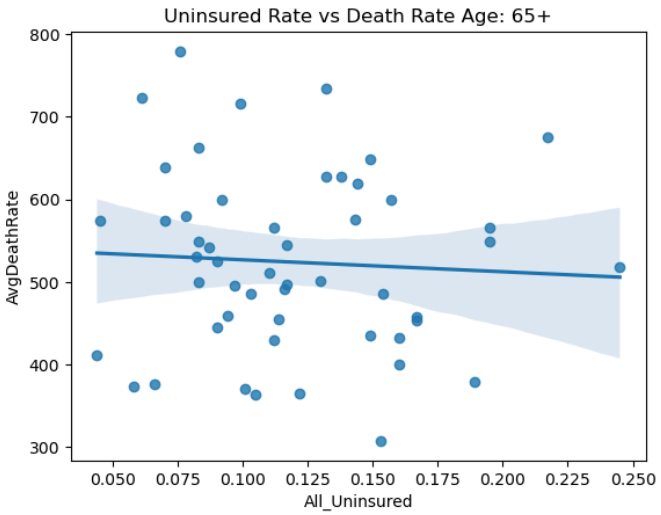
Code

```
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()
```



Code

```
sns.regplot(data=state_df_65, x='All_Uninsured', y='AvgDeathRate', scatter=True)

plt.title('Uninsured Rate vs Death Rate Age: 65+')

plt.show()
```

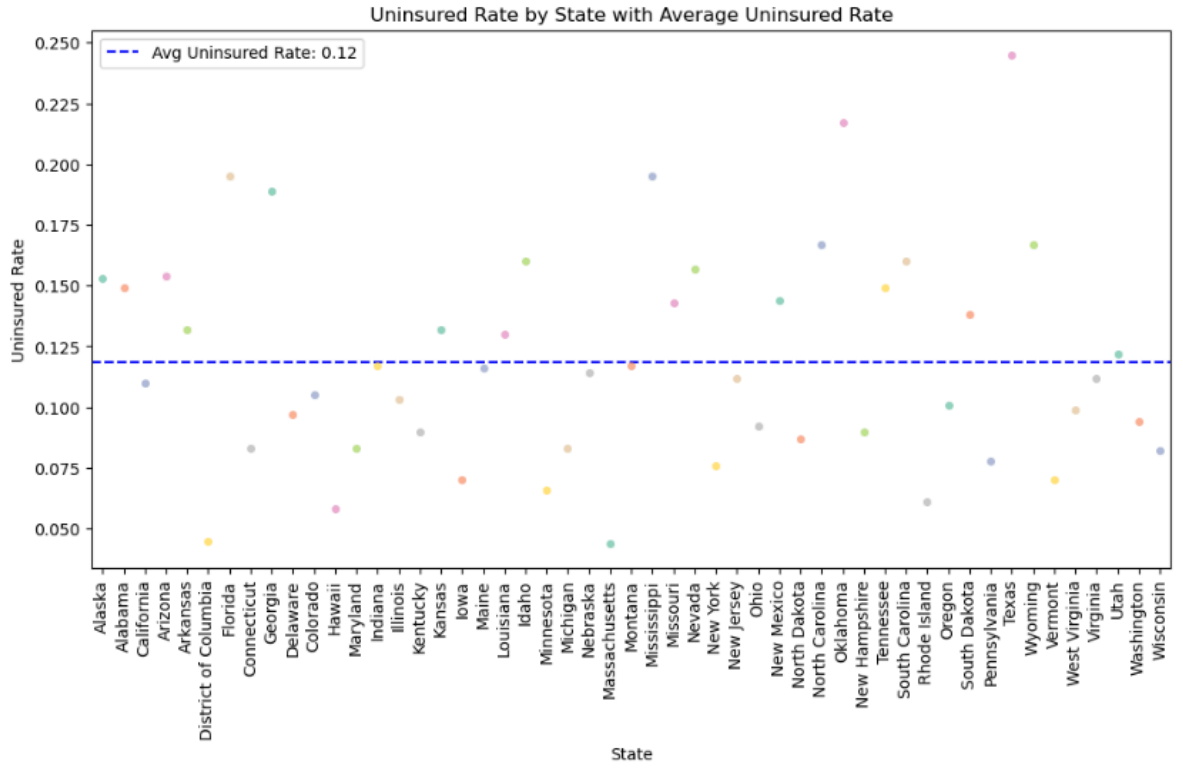


For the 0–44 and 45–64 age groups, the regression plots show a clear positive relationship, with the best-fit line indicating that the uninsured rate has predictive power for death rate. In contrast, for the 65+ age group, the scatter plot lacks a clear trend, and the best-fit line has a shallow slope, suggesting that the uninsured rate has limited predictive power for death rate. However, we will explore incorporating state-level factors and assess how a more complex model, such as SVR, performs.

Below is the code to visualize the uninsured rate across different states. This visualization allows us to easily compare each state's uninsured rate, highlighting variations and trends between states.

Code

```
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()
```



5. SQL Script and Schema

Script

The file knm_datasetup.sql contains the SQL script to load data into the database.

Schema

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)

USCDI_cancer(LocationDesc, Cancer_Deaths, Cancer_Deaths_F, Cancer_Deaths_M, CancerPercentage, CancerPercentage_F, CancerPercentage_M)

state_df(LocationDesc, DeathRateUnit, DeathRateType, AvgDeathRate, Stratification1, All_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
-