Investigating the Relationship Between Heart Disease Mortality and Obesity Rates of Adults in the United States

STAT 447 Final Project Report

Katie Chai (kchai2), Josh Placko (jplacko2), Michael Hu (mhu18), Rishi Mohan (rishinm2)

December 7, 2022

Contents

1 Introduction

In the United States, heart disease is the leading cause of death for adults and obesity is considered an epidemic. As two of the most important health problems nationwide, heart disease mortality and obesity are considered to be closely related to each other. The goal of this project was to investigate the extent of their relationship, as well as how this relationship is affected by different demographics of people, such as gender, ethnicity, age, etc. Another goal was to seek out any unique trends or findings within the data throughout the investigation.

2 Data

The data used in the project was obtained from metadata published by the United States' Centers for Disease Control and Prevention (CDC). Two different datasets, "Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County - 2018-2020" and "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System", were downloaded as .csv files from the CDC data website, and used concurrently in the analysis.

The Heart Disease Mortality dataset contains 3-year averaged values of heart disease mortality rates per 100,000 people for different stratifications on national and state levels.

The Nutrition, Physical Activity, and Obesity dataset contains yearly data related to the diet, physical activity, and weight status of adults in the United States. There is extensive data, including that for different years, counties, questions, but only a small subset that corresponded to the limitations placed by the Heart Disease Mortality data was used. This limitation will be further discussed in the 'Data Wrangling' section of the report.

3 Setup

R via RStudio was used as the main setup in this project, from data wrangling to investigative graphing and modeling. Version control using Git and GitHub was also implemented as a means of collaboration between group members.

4 Methods

4.1 Data Wrangling

The raw data downloaded from the CDC website was initially in a form that was very difficult for analysis, making proper data wrangling crucial for later aspects of the project. Wrangling of the data was performed using the following R packages: `dplyr`, `data.table`, `readr`, `tidyr`.

4.1.1 Heart Disease Mortality Data

This dataset contained only national and state level data, and a total of three stratifications (overall, gender, and ethnicity). The raw dataset was filtered for state level data only, and split by its three stratifications for modeling analysis with the Nutrition, Physical Activity, and Obesity dataset. The R script for wrangling of the Heart Disease Mortality data is available in the **split_heart_data_script.R** file within **/data/analysisdata/**.

4.1.2 Nutrition, Physical Activity, and Obesity Data

This dataset contained data on the national, state, and county levels, as well as six different stratifications (overall, gender, ethnicity, age, education, and income level). However, due to the limited data in the Heart Disease Mortality dataset, only the corresponding subsets (state level data stratified overall, to gender, and to ethnicity) could be used in modeling.

Furthermore, it also had extensive data regarding different aspects of nutrition, physical activity, and obesity of American adults. For example, there were questions regarding the percentage of adults who report consuming fruits or vegetables a certain number of times daily, or others regarding the percentage of adults who engage in physical activity a certain number of times weekly. The responses to these questions are, however, all very closely related to obesity rates, which is why only the question regarding the percentage of adults who have obesity was used in the main analysis. The R script for wrangling of the Nutrition, Physical Activity, and Obesity data is available in the **split_nutr_data_script.R** file within **/data/analysisdata/**.

4.1.3 Combined Data

After filtering and splitting the two datasets, the corresponding subsets were combined to have data for both obesity and heart disease mortality rates at the state level, stratified by overall, gender, and ethnicity. The R script for combining the two datasets is available in the **sub_nutr_data_script.R** file within **/data/analysisdata/**.

4.2 Visualizations

Two main types of visualizations were created: scatter plots of heart disease mortality rates versus obesity rates to visualize the general trend between the two variables based on different stratifications, and choropleth maps to visualize the geographic differences in the same variables. Both graph types utilized the `ggplot2` package of R.

4.2.1 Scatter Plots

The scatter plots were made simply by plotting the heart disease mortality rates values on the y-axis versus obesity percentage values on the x-axis. Each individual data point represented a US state/territory, and if stratified, the points were colored or plots separated by each category within the stratification.

4.2.2 Choropleth Maps

The choropleth maps were made by filling each respective state with either the heart disease mortality rate or obesity rate value. If stratified, the plots were separated by each category within the stratification.

4.3 Modeling

The models for this project were fit linearly using the `lm` function, between heart disease mortality rate and obesity rate, along with the stratification categories as factor variables.

4.4 Shiny App

A Shiny app was created as a way to effectively display all results in an interactive manner. General functions were created from the code developed in previous steps, allowing for the reactive use of visualizations and model summaries. The code for these functions are available in files **graph_function.R** and **mapvis_function.R** within **/graphs/** and **/mapvisualization/**, respectively.

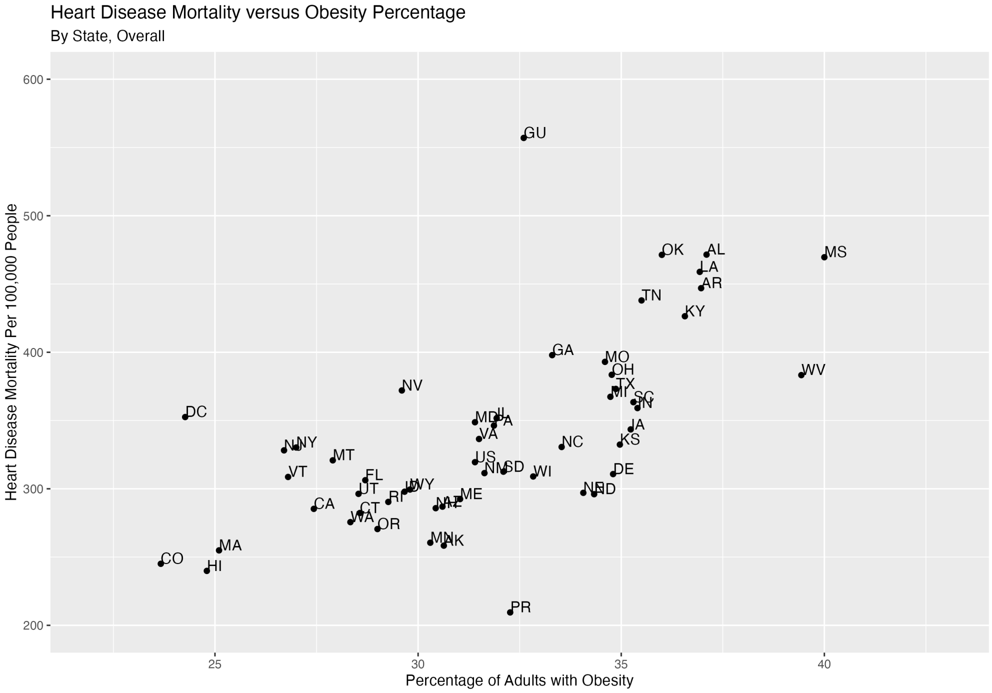
Development of the Shiny app was done using the following R packages: `shiny`, `shinydashboard`, `shinyWidgets`, and `plotly`. The entire code needed to recreate the app is available in the **app.R** file of the main directory, and a working version of the app is available at this link.

5 Results

5.1 Visualizations

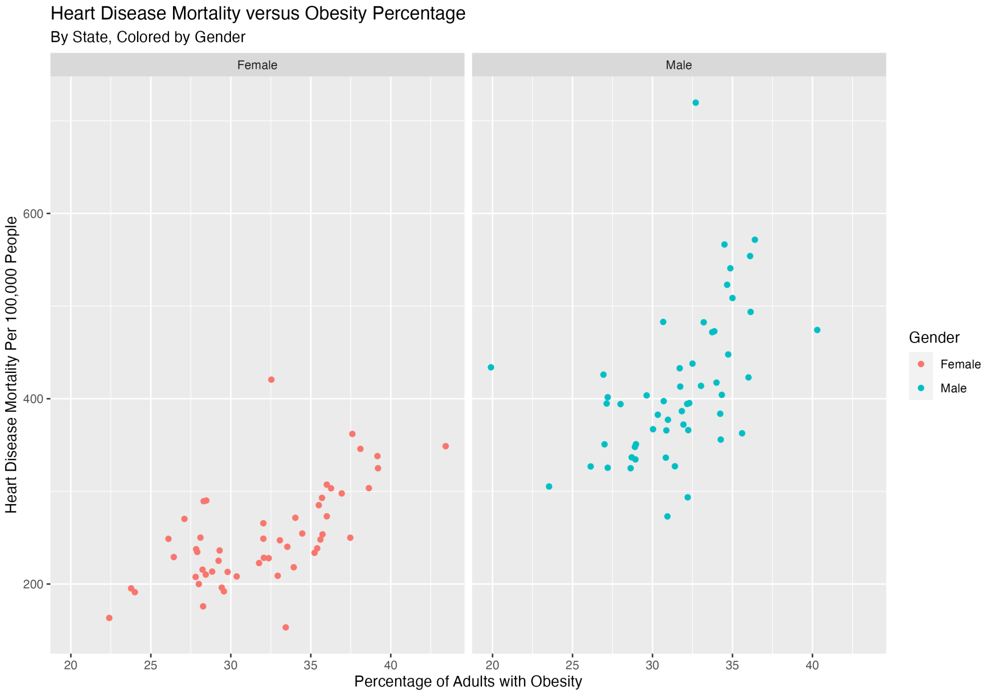
5.1.1 Scatter Plots

5.1.1.1 Overall Trend Between Heart Disease Mortality and Obesity Rates



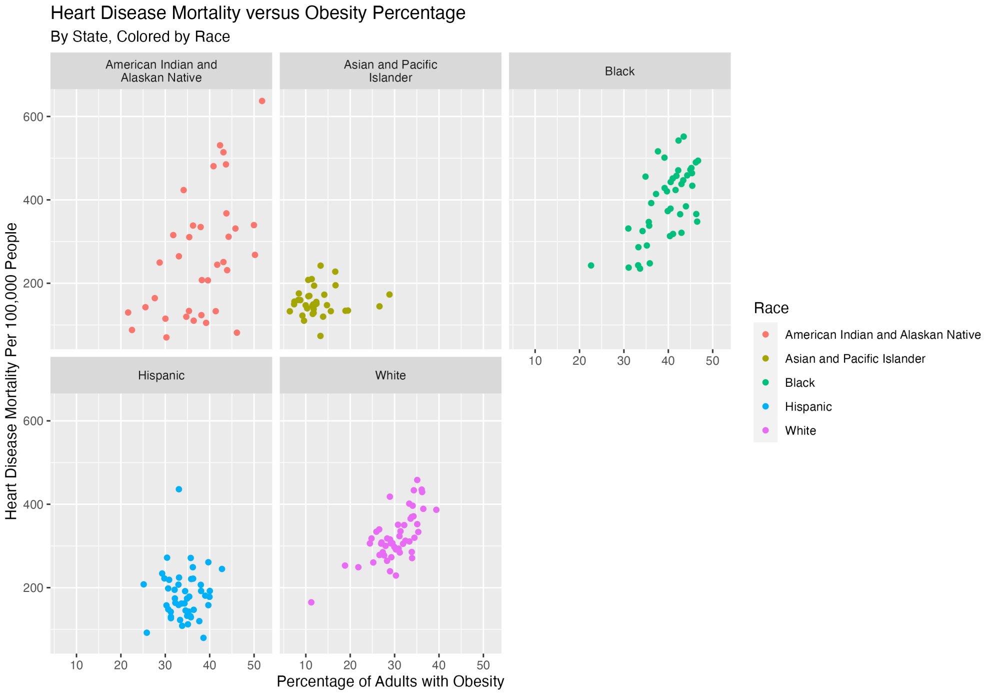
Based on this graph, it is evident that there is a generally increasing trend between the two variables. An outlier to note, however, is Guam (GU). While most of the US states and territories lie within a similar range of values, Guam has a much higher heart disease mortality for corresponding obesity rates. A possible explanation for this could be the fact that it is located the furthest away from mainland America out of all the US states and territories, which could result in difficulties with obtaining preventative medical resources.

5.1.1.2 Trends Separated by Gender



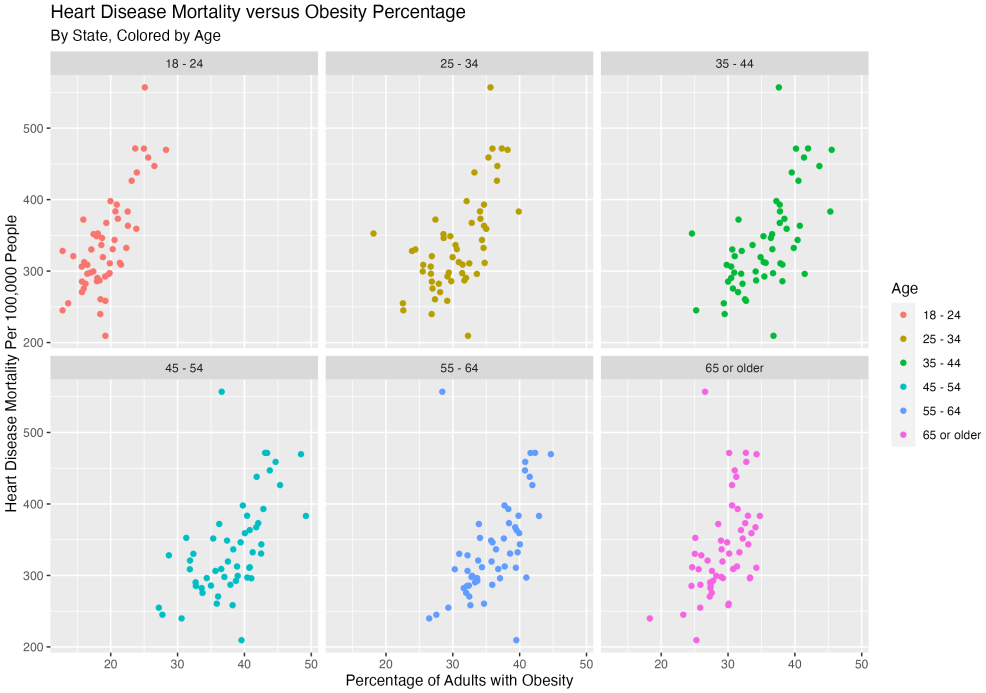
Based on these graphs, it is evident that while both genders have a similar distribution of obesity rates, the women seem to have much lower heart disease mortality rates compared to that of men. A possible explanation for this result could be that women are more resilient to heart disease than men, but further investigation is required to determine the feasibility of this interpretation.

5.1.1.3 Trends Separated by Ethnicity



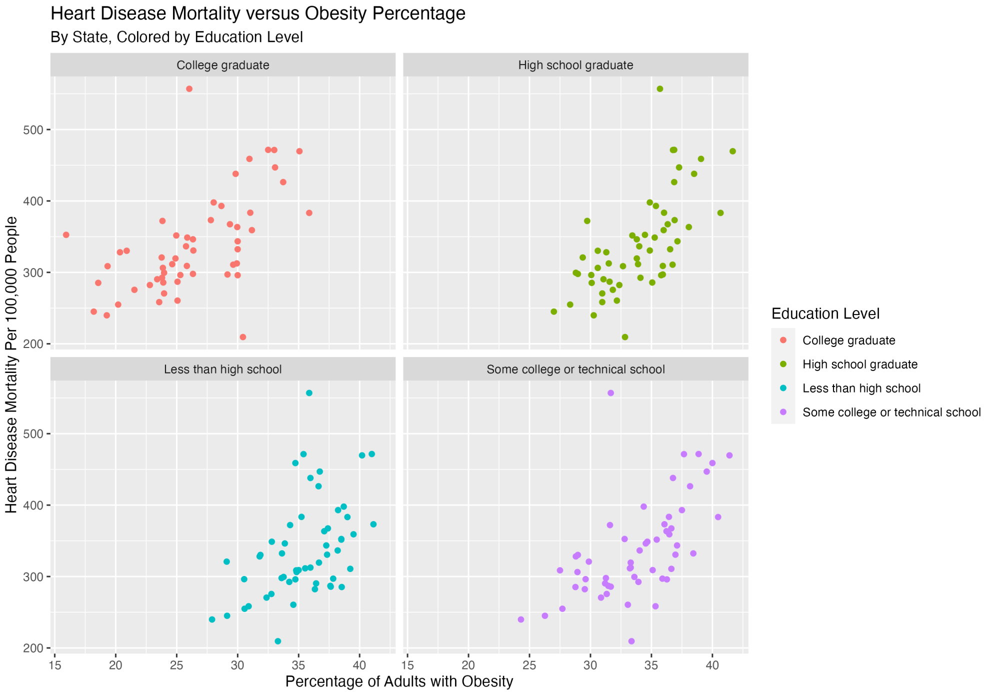
From these graphs, it can be seen that for the American Indian and Alaskan Native population, there is a much greater spread of heart disease mortality rates compared to other ethnicities, while for the Asian and Pacific Islander (API) and Hispanic populations, there is a much smaller spread of both heart disease mortality and obesity rates, with the API population having lower values for both variables, and the Hispanic population having lower mortality rates but relatively higher obesity rates. The Black and White populations both have higher rates for mortality, with the Black population having higher rates of obesity as well, while the White population has mid-range obesity rates. Possible reasons for these dissimilarities could arise from genetic or cultural differences that lead to healthier (or unhealthier) lifestyles.

5.1.1.4 Trends Separated by Age Group



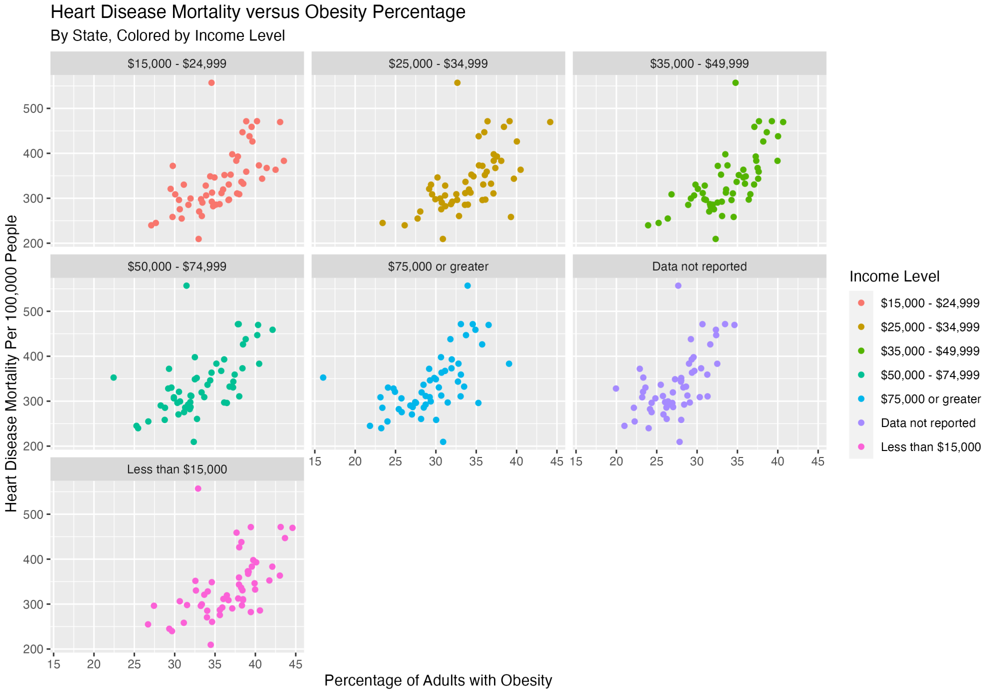
These graphs show that for the same values of heart disease mortality, the 18-24 age range has the lowest range of obesity rates, with these rates increasing with age until the 55-64 age range, at which they start to decrease again. These trends can be intuitively understood with the fact that younger people have higher metabolisms and also tend to be the more active. The decrease in obesity rates for the 55-64 and 65 and older age ranges could possibly be explained by the idea that many of the obese persons at this age actually passed away, either due to heart disease or other health issues, compared to their non-obese counterparts.

5.1.1.5 Trends Separated by Education



It can be observed based on these graphs that college graduates seem to have much lower rates of obesity compared to those with less education (those who completed less than high school, high school graduates, or those who completed some college or technical school). A possible explanation for this could be that college graduates are more educated on healthier lifestyles, or that those who can afford college would also be able to afford a healthier lifestyle.

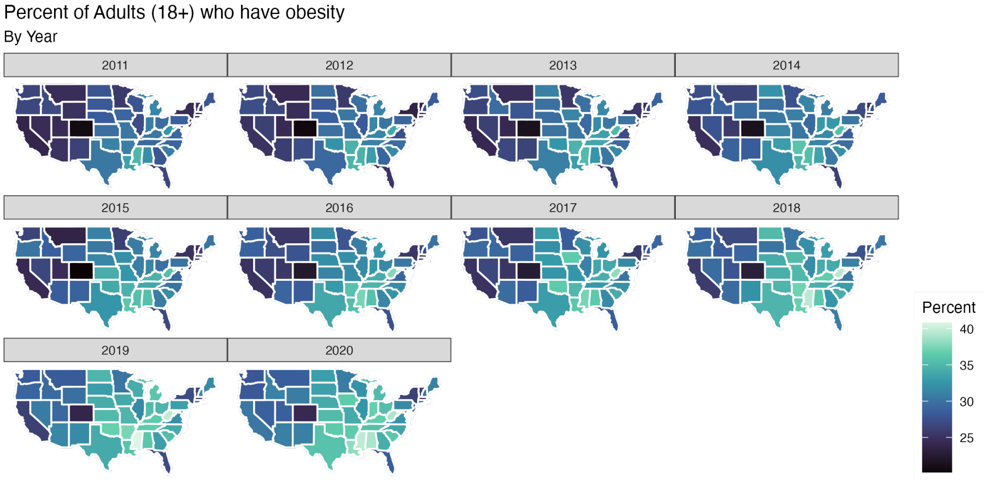
5.1.1.6 Trends Separated by Income Level



The graphs show that the income group earning less than $15,000 has the highest rates of obesity, and obesity tends to decrease as income level increases. These findings could possibly be explained by the idea that with more money, households are able to afford more resources for the improvement of their health. Interestingly, the group with no data reported has the lowest rates of obesity. If following the observed trend of lower obesity for higher income, it could be that households that make significantly more than the maximum income level of $75,000 did not want to report their income. However, further investigation is needed to determine the reality of these trends.

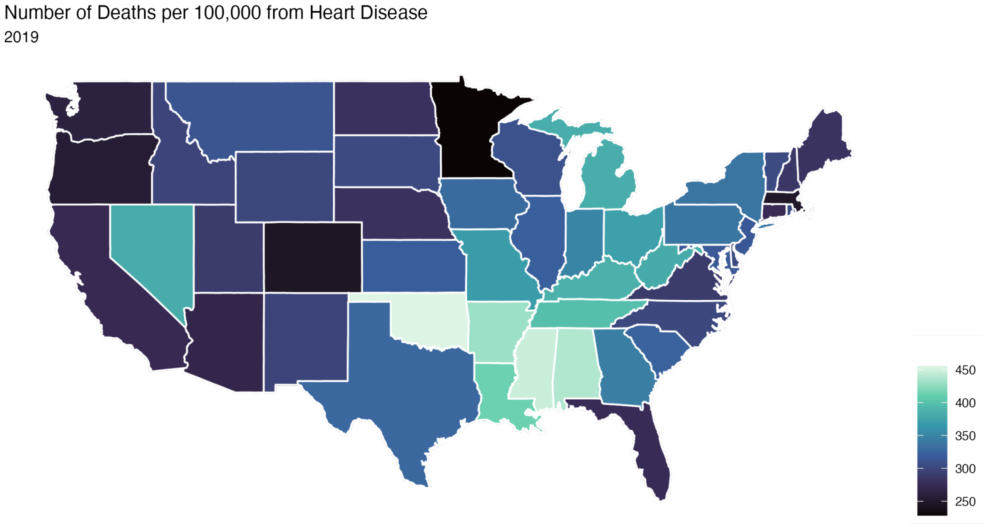
5.1.2 Choropleth Maps

5.1.2.1 Obesity Rates by Year and State



This visualization shows that there is a consistent increase in obesity rates for almost all states over time. It also shows that the southern and central parts of the US tend to have higher obesity rates compared to the coastal areas. Interestingly, Colorado (CO) remains consistently low in its obesity rates throughout the years, which could possibly be explained by an increased 'outdoor' culture, but this again requires further investigation for justification.

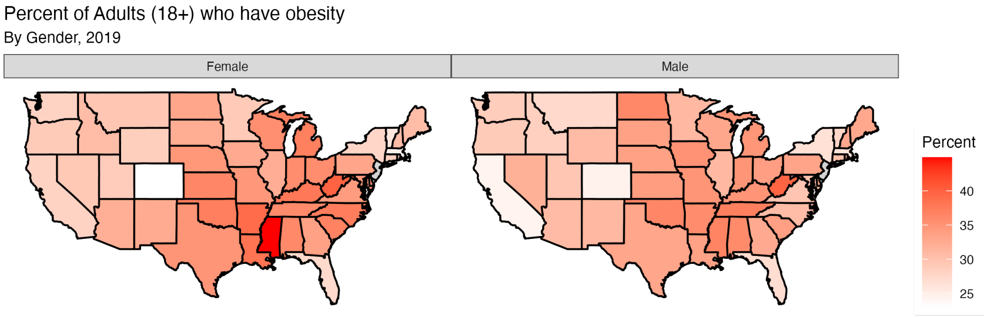
5.1.2.2 Overall Heart Disease Mortality Rates by State


Number of Deaths per 100,000 from Heart Disease

2019

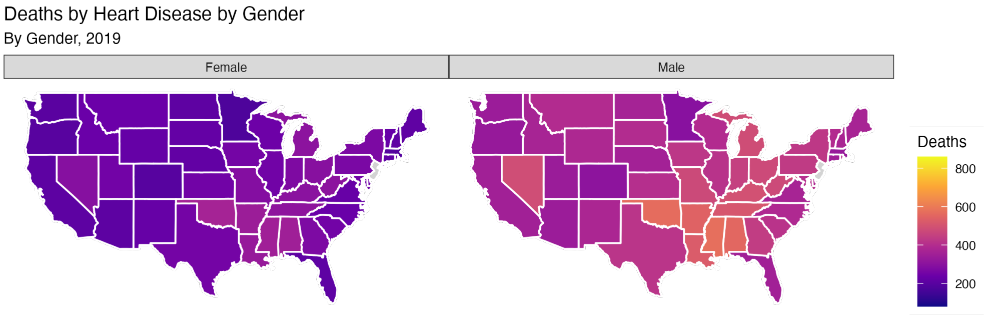
This visualization shows that the southern and central parts of the US also tend to have higher heart disease mortality rates. Interestingly, Nevada (NV) has a significantly high mortality rate despite having low obesity rates. Further investigation could provide insight into why this is.

5.1.2.3 Gender-Stratified Obesity Rates by State



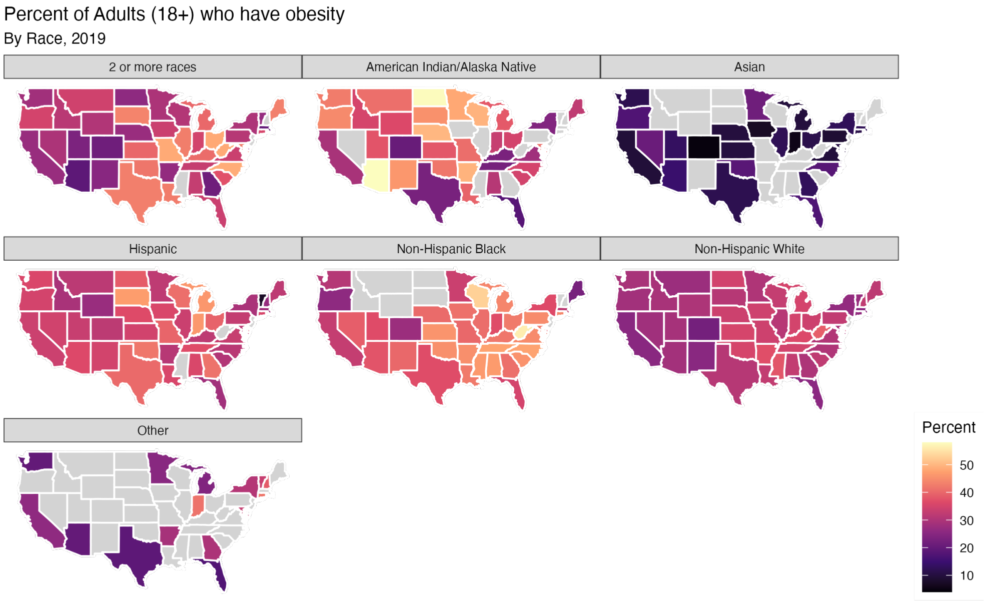
This visualization shows that despite the gender-stratified general graph showing that men and women had a similar distribution of obesity rates, California (CA) and Mississippi (MS) are two states where the women have noticeably higher obesity rates than men.

5.1.2.4 Gender-Stratified Heart Disease Mortality Rates by State



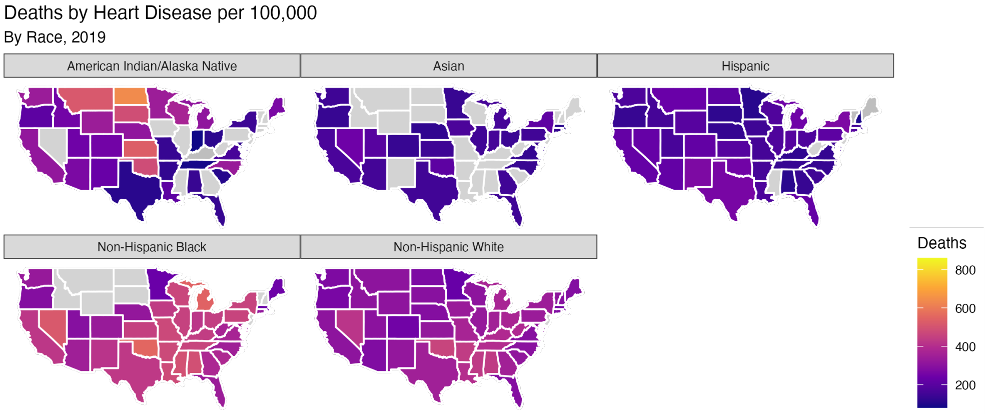
This visualization shows that as seen in the gender-stratified general graph, men noticeably tend to have higher rates of mortality due to heart disease than women for almost all states.

5.1.2.5 Ethnicity-Stratified Obesity Rates by State



This visualization, while having many missing values, shows that generally, Asian populations have significantly lowest obesity rates, and American Indian and Alaskan Native, Hispanic, and non-Hispanic Black populations having the highest obesity rates. Interesting results to note are the significantly high rates of obesity for American Indian populations in Arizona (AZ) and North Dakota (ND).

5.1.2.6 Ethnicity-Stratified Heart Disease Mortality Rates by State



This visualization, again, while having many missing values, shows that Asian and Hispanic populations have the lowest rates of heart disease mortality, while American Indian and Non-Hispanic Black populations have the highest rates. Again, North Dakota (ND) has significantly high mortality rates for the American Indian population.

5.2 Modeling

5.2.1 Overall Linear Model

The coefficients from the linear model fitting the overall relationship between obesity and heart disease mortality rate are shown below.

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -32.71363  61.292699 -0.533728 5.958040e-01
obese_perc   11.66796   1.915179  6.092359 1.375746e-07
```

The p-value for the `obese_perc` slope estimate is much less than 0.05, indicating that at a significance level of 0.05, the effect of obesity rate on heart disease mortality rate is statistically significant. The coefficient estimate of 11.67 represents that for a 1% increase in obesity rate, there will be 11.67 more deaths per 100,000 in heart disease mortality.

5.2.2 Gender-Stratified Linear Model

The coefficients from the linear model fitting the gender-stratified relationship between obesity and heart disease mortality rate are shown below.

```
                              Estimate Std. Error    t value     Pr(>|t|)
(Intercept)                 -33.026929  46.749049 -0.7064727 4.814893e-01
obese_perc                    8.868967   1.436769  6.1728568 1.342817e-08
factor(Stratification1)Male 165.817485  11.605396 14.2879644 3.499198e-26
```

For this model, the p-values for both the `obese_perc` and `factor(Stratification1)Male` estimates are much less than 0.05, indicating that the effect of obesity rate on heart disease mortality rate is statistically significant even after considering gender differences in both variables, and that the gender difference in mortality is also statistically significant. The slope estimate of 8.87 represents that for a 1% increase in obesity rate, there will be 8.87 more deaths per 100,000 in heart disease mortality, and the coefficient estimate for the male group of 165.82 represents that at 0% obesity, men have 165.82 more heart disease-related deaths than women. Since the model is a main-effects-only model, the results indicate that men and women have the same effect of obesity on heart disease mortality.

5.2.3 Ethnicity-Stratified Linear Model

The coefficients from the linear model fitting the ethnicity-stratified relationship between obesity and heart disease mortality rate are shown below.

```
                                               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)                                  -19.586920 39.1046880 -0.5008842 6.169848e-01
obese_perc                                     7.412155  0.9767859  7.5883107 1.084556e-12
factor(race_label)Asian and Pacific Islander  80.259220 30.1598295  2.6611298 8.398713e-03
factor(race_label)Black                      119.281274 16.8213418  7.0910677 2.065819e-11
factor(race_label)Hispanic                   -54.246441 16.7290884 -3.2426418 1.380440e-03
factor(race_label)White                      117.327731 17.8336292  6.5790160 3.804178e-10
```

For this model, the p-values for all coefficient estimates excluding the intercept are much less than 0.05, indicating that the relationship of obesity on heart disease mortality remains significant after consider racial differences in mortality, and that each of the ethnicities have statistically significant differences in mortality rates as well. The slope estimate of 7.41 represents that for a 1% increase in obesity rate, there will be 7.41 more deaths per 100,000 in heart disease mortality. The coefficient estimates for the API, Black, Hispanic, and White groups are 80.26, 119.28, -54.25, 117.33, respectively. These represent the differences in heart disease related deaths per 100,000 at 0% obesity of the respective races compared to the American Indian population. Again, since the model is a main-effects-only model, the results indicate that all races have the same effect of obesity on heart disease mortality.

5.2.4 Question-Stratified Linear Model

As aforementioned in the 'Nutrition, Physical Activity, and Obesity Data' section, the original dataset contained extensive information on many different questions regarding the nutrition, physical activity, and obesity of American adults. A model investigating the difference in heart disease mortality rates based on the nutrition, physical activity, or obesity related question answered was fit, with the results shown below.

```
                        Estimate Std. Error    t value     Pr(>|t|)
(Intercept)           350.819231 30.2326308 11.6039928 1.505730e-27
```

```
obese_perc               -0.311818  0.7127003 -0.4375163 6.619380e-01
factor(QuestionID)Q019  -6.113986 19.4395480 -0.3145128 7.532712e-01
factor(QuestionID)Q036  -2.827343 14.7421861 -0.1917859 8.479927e-01
factor(QuestionID)Q037  -1.878416 13.9950810 -0.1342197 8.932863e-01
factor(QuestionID)Q043   3.352338 15.5346109  0.2157980 8.292388e-01
factor(QuestionID)Q044  -5.572717 18.5701425 -0.3000902 7.642411e-01
factor(QuestionID)Q045  -2.509841 14.6807014 -0.1709619 8.643273e-01
factor(QuestionID)Q046  -1.753829 14.0955210 -0.1244246 9.010323e-01
factor(QuestionID)Q047  -4.973209 17.3371338 -0.2868530 7.743513e-01
```

Looking at the p-values of the coefficients, it can be determined that there was no statistically significant difference on the effect of obesity rate on heart disease mortality for each of the different questions. Again, as aforementioned in the obesity data wrangling section, this can be explained by the interconnectedness between the results of each of these questions.

6 Conclusions and Suggestions

In conclusion, the data graphs and modeling results showed evidence of a strong increasing trend between rates of obesity and heart disease mortality. Males, and the Black, White and API populations, were found to have a higher rate of heart disease mortality at the same obesity levels compared to their female, and Hispanic and American Indian, counterparts.

Based on the results of the choropleth maps, many questions for further investigation arose. For one, why are the rates of both heart disease mortality and obesity much higher in the southern and central states? Why does Nevada have a noticeably high rate of heart disease related deaths despite having lower obesity rates? Why does Colorado consistently have lower obesity rates compared to the rest of the country? Why are obesity rates for women significantly higher in California and Mississippi and for American Indian populations in Nevada and North Dakota? What makes these states unique? In order to answer these questions, more data must be collected.

It also could be of interest to develop different types of models. Only main effects linear models were used in this investigation, which are very limited. Intuitively, the inclusion of interaction terms could be necessary, since it is more than likely that the stratifications affect both obesity rate and heart disease mortality rate at the same time, meaning that different stratification categories likely have different relationships between obesity and heart disease mortality rates.

Overall, to perform any further analyses, the collection of more data is necessary. For the heart disease mortality data especially, gathering data based on more stratifications and geographic levels could be helpful in further investigative attempts.