DataHacks2020

Intro

America has been known for its high obesity rate; in this report, we will explore US health data to identify some potential causes for obesity and the repercussions of it. The data that we are exploring comes from the US Centers for Disease Control and Prevention and contains state and national statistics for various topics and queries. We will be focusing on obesity prevalence, diabetes mortality rates, and cardiovascular disease mortality rates.

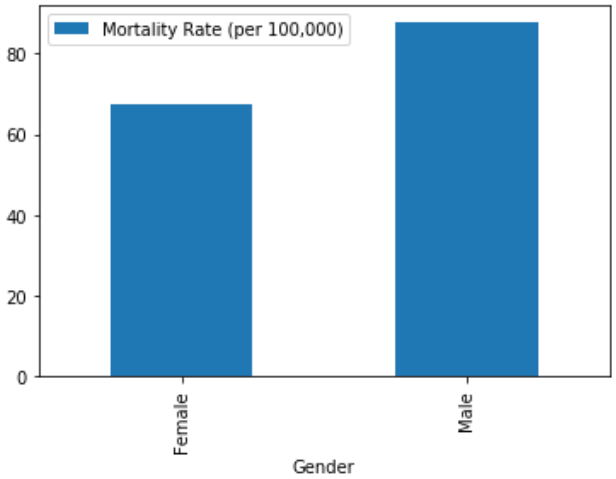
Data Cleaning and Pre-Processing

Upon examining the data, we notice that there are many columns (10) that consist entirely of either one value or null values. We can safely remove those first. The TopicID column abbreviates the Topic column; since Topic is more specific, we can remove TopicID because it is redundant. It is worth noting that the column DataValue contains null values and is an object, and not a float like one would expect. This is because it contains blank values ' ', 'Yes', and 'No' values. We won't examine any Boolean columns, so we only have to remove the blanks values.

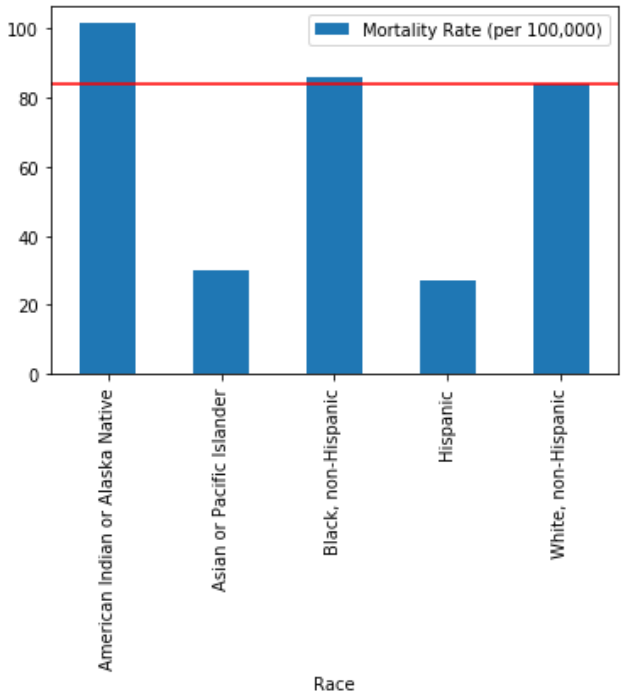
Visualizations

When looking at the data, we will find that there is the most data on diabetes, so we will begin our exploration of the data there. To get a sense of how diabetes affects different demographics, we plot the diabetes mortality rate first by gender, then by race:

Obesity Mortality by Gender

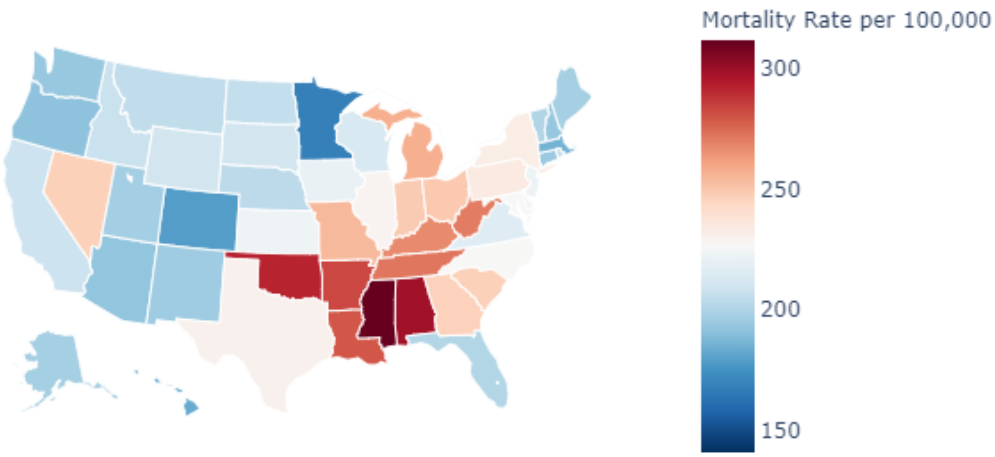


Obesity Mortality by Race

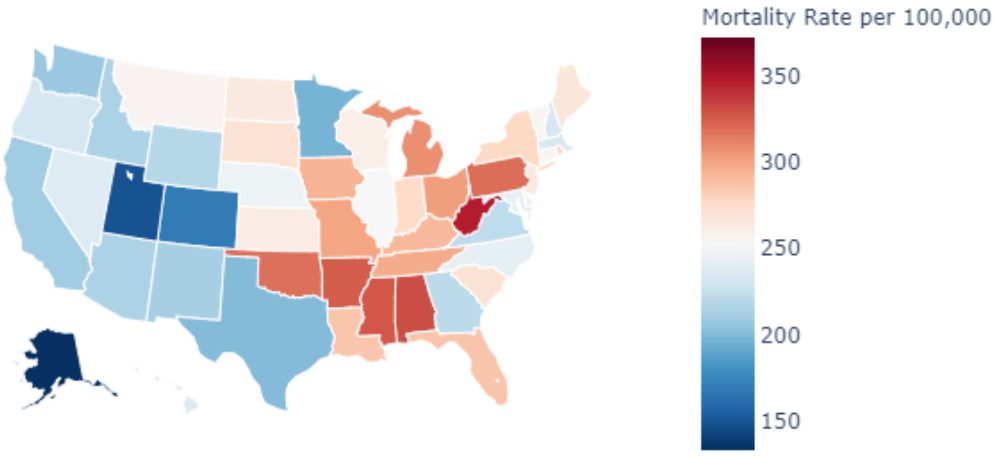


Next, we looked at diabetes mortality rates by state. For each year, each state has two mortality rates: the crude rate and the age adjusted weight. For older populations, the prevalence of diabetes and other health complications increases, so the age adjusted weights lower prevalence for populations with a high average age. We plotted both types of rates to compare:

Adjusted Average Cardiovascular Disease Mortality Rate by State (2010-2014
US Average: 226.24



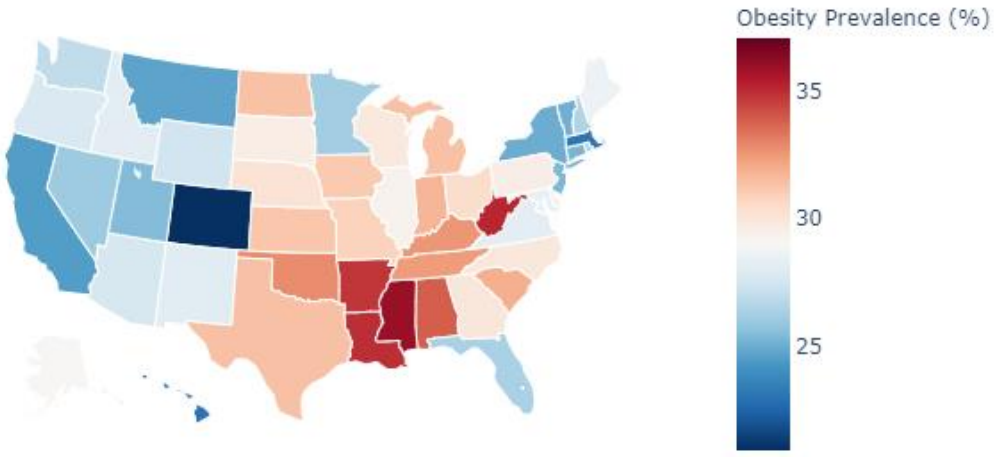
Crude Average Cardiovascular Disease Mortality Rate by State (2010-2014)
US Average: 252.6



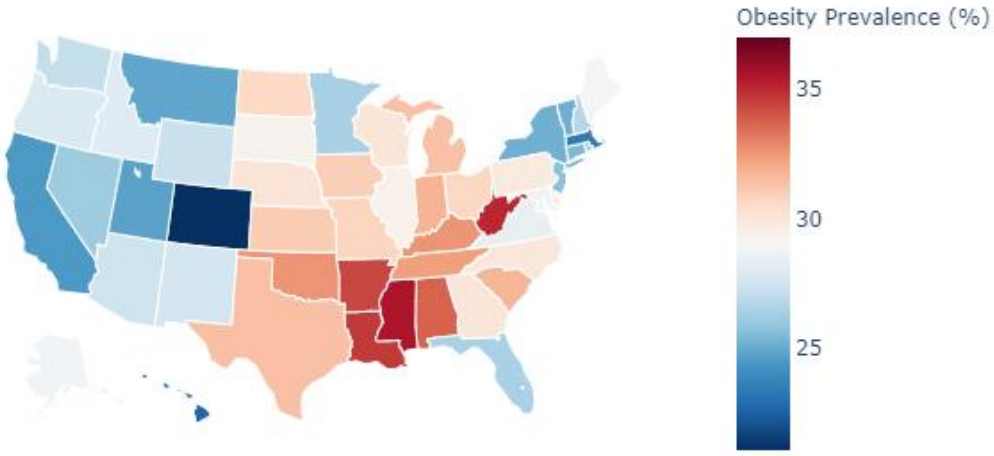
From these plots, we say that West Virginia and Oklahoma have concerningly high diabetes mortality rates. From prior knowledge, we know that obesity and inactive lifestyles can contribute to develop type 2 diabetes. We can then hypothesize that these states will also have high obesity prevalence. One of the topics, 'Nutrition, Physical Activity, and Weight Status,' covers obesity rates.

Using a similar process, we found the obesity prevalence of each state (both crude and age adjusted):

Adjusted Obesity Prevalence by State (2011-2016)
US Obesity Prevalence: 28.98%

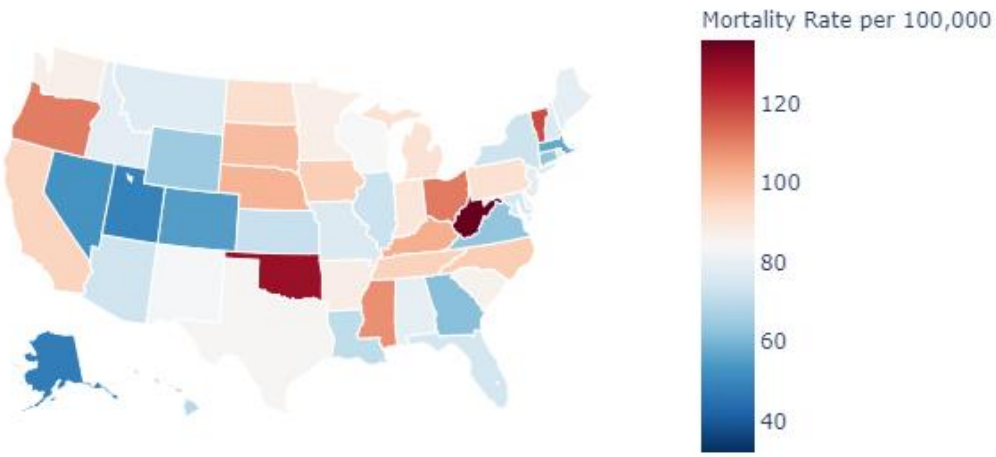


Crude Obesity Prevalence by State (2011-2016)
US Obesity Prevalence: 29.02%

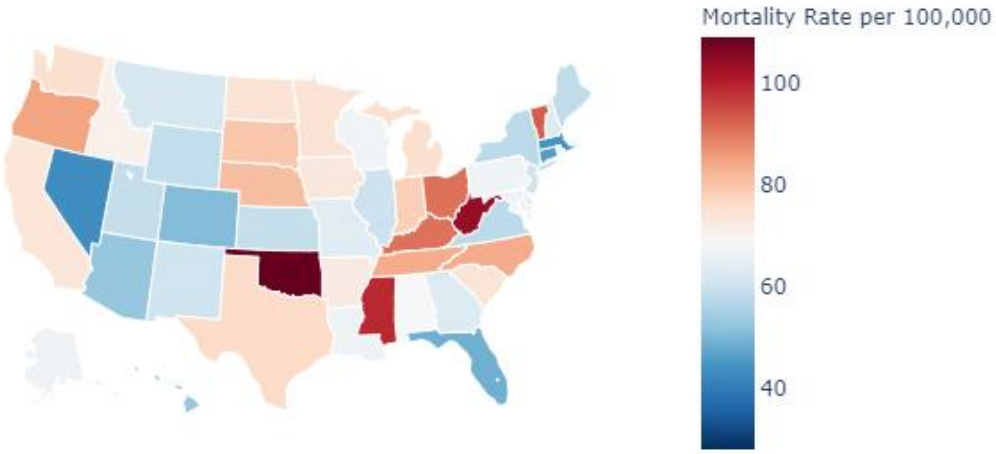


As we hypothesized, these states have rather high obesity prevalence (especially West Virginia). We know that obesity increases the risk of cardiovascular disease, so we examined cardiovascular disease mortality next:

Crude Average Diabetes Mortality Rate by State (2010-2014)
US Average: 83.94



Adjusted Average Diabetes Mortality Rate by State (2010-2014)
US Average: 68.34



As expected once again, these states have high cardiovascular disease mortality.

These plots were produced by Plotly's graph object choropleth map. In these charts, white represents the national average; blue colors represent below average mortality rates, while red colors represent above average mortality rates. In a notebook, hovering over a state will display the state's mortality rate.

Analysis

The extent of the natural language processing in this project is rather limited. To clean the data, we stripped strings and converted them to floats for plotting.

We also ran a permutation test to determine if the population of West Virginia's obesity prevalence came from the same distribution as the population of the US's obesity prevalence, and we set a significance level of 0.01. We then assumed that the data was gathered by sampling ~1000 individuals, so we generated 50,000 samples of 1000 individuals who had a ~29% (the national average) of being obese. We ended with a p-value of ~0.00, so we rejected the null hypothesis, meaning that the population of West Virginia is different from the national population.

Proposal

From the visualizations, we saw that certain states, especially West Virginia and Oklahoma, have high obesity rates. This likely lead to high rates of other health complications, such as cardiovascular disease and diabetes. We should then study what causes the obesity prevalence to be so high by examining local lifestyles, cuisine, and genetics to better understand how to reduce these health issues.