Heart Disease in the United States

Richardson, Trevor K./ Liu, Chi-Yun

2020-12-06

Introduction

In the United States, heart disease is one of the leading causes of death with 655,000 deaths per year costing the US about $219 billion per year (Center for Disease Control and Prevention [CDC], 2020). Essentially, we were interested in analyzing data gathered by the Center for Disease Control and Prevention (CDC) and determine how it affects Americans by varying variables such as race and gender. Upon initial look at the dataset, there are numerous variables of interest: race/ethnicity, state, county and rates per 100,000. This allowed us to come up with a few hypothesis on how heart disease affects Americans by varying factors.

Initial Hypotheses

- **Hypothesis 1**: Heart disease is higher in the southern states of the United States.

Since the dataset consists of geographical variables, it was of interest to examine if rates of heart disease differed by region. There are varying regions in the United States such as the Midwest, North Atlantic, Pacific Northwest, New England and Southern States. Since populations typically differ in these varying regions, we can assume that heart disease affects

regions differently. We hypothesize that the rates of heart disease is higher in the southern states of the US, most commonly referred to "The South", and as stated, this can be due to the demographic makeup of the varying regions.

- **Hypothesis 2**: Heart disease is higher in African Americans

The dataset included a race/ethnicity variable which can allow us to tidy the data by race/ethnicity. It can be assumed that since the CDC included this variable in the collection of the data that it can be examined further. It is commonly known that race and ethnicity can be an attributing factor in many issues in American society. As such, we can assume heart disease affects each race/ethnicity differently, and we hypothesize that African Americans have higher rates of heart disease.

- **Hypothesis 3**: Heart disease is higher in American males

The dataset included a gender variable and this can allow us to tidy the data by gender. Since the CDC included this variable in the collection of the data, it would be of interest to see what the rates of heart disease are between males and females. Since it is commonly known that there are differences between the two sexes, we hypothesize that males have higher rates of heart disease than females.

Exploratory Data Analysis

The dataset was obtained from the CDCs official website and can be seen with the following link: <https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/i2vk-mgdh> The dataset contains data gathered in adults ages 35 and older from 2013 to 2015. There are 19 columns (Variables) and 59076 rows (Observations) of data and the variables are shown below:

```
## [1] 59076    19
```

1. Year - numeric data, center of 3-year average
2. LocationAbbr - character data,State, Territory, or US postal abbreviation
3. LocationDesc - character data, counties
4. GeographicLevel - character data, geographic level
5. DataSource - character data, not needed
6. Class - character data, not needed
7. Topic - character data, not needed
8. Data_Value - numeric data, heart disease mortality rate (rates per 100,000)
9. Data_Value_Unit - character data, not needed
10. Data_Value_Type - character data, not needed
11. Data_Value_Footnote_Symbol - character data, not needed
12. Data_Value_Footnote - character data, not needed
13. StratificationCategory1 -character data, not needed
14. Stratification1 - character data, gender categories
15. StratificationCategory2 - character data, not needed
16. Stratification2 - character data, race/ethnicity categories
17. TopicID - character data, not needed
18. LocationID - character data, not needed
19. Location 1 - character data, latitude and longitude

There are quite a bit of variables that are not useful for our research and some variables that contain mainly NA values, so after tidying the data, we are left with 9 variables that we will manipulate for our research as seen below:

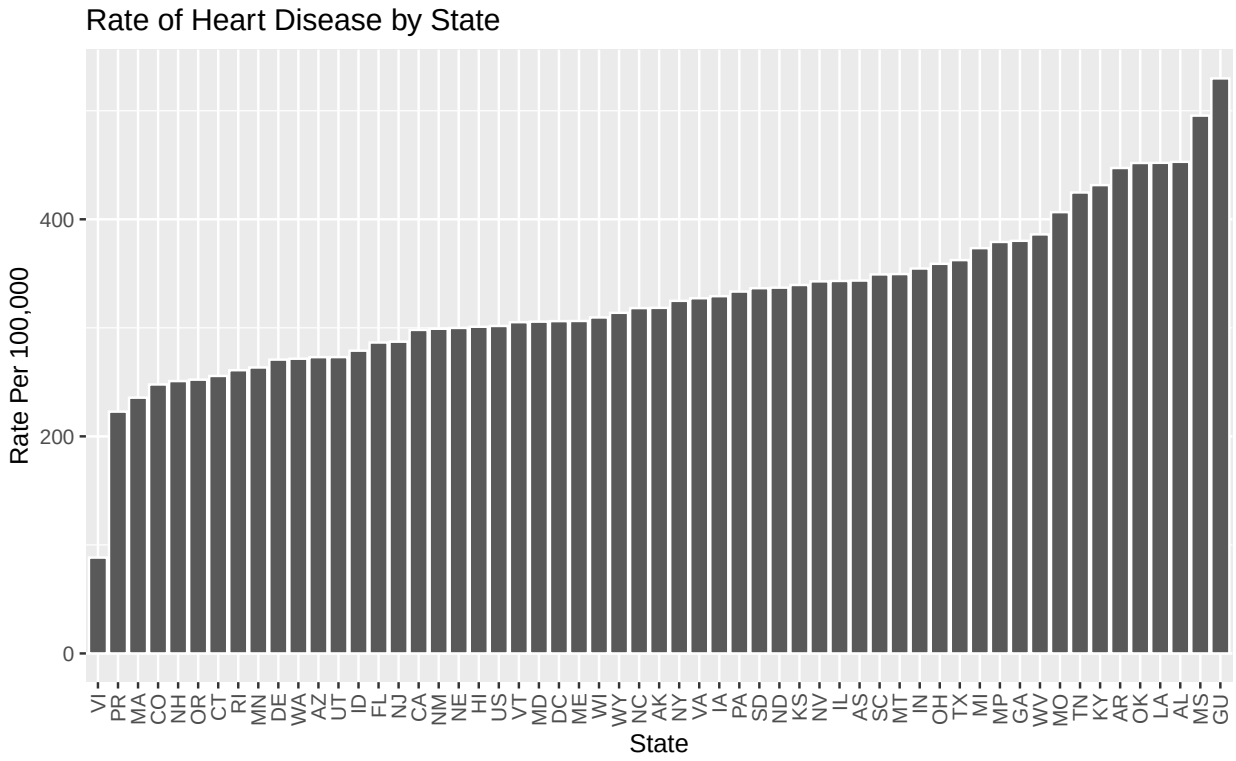
```
## [1] "City"           "Gender"          "GeographicLevel" "LocationLat"
## [5] "LocationLon"    "RaceEthnicity"   "RatePer100Thous" "State"
## [9] "Year"
```

The tidied dataset will allow us to focus on the variables: state, city, heart disease mortality rate, and gender and this will allow us to investigate if the data supports our hypotheses.

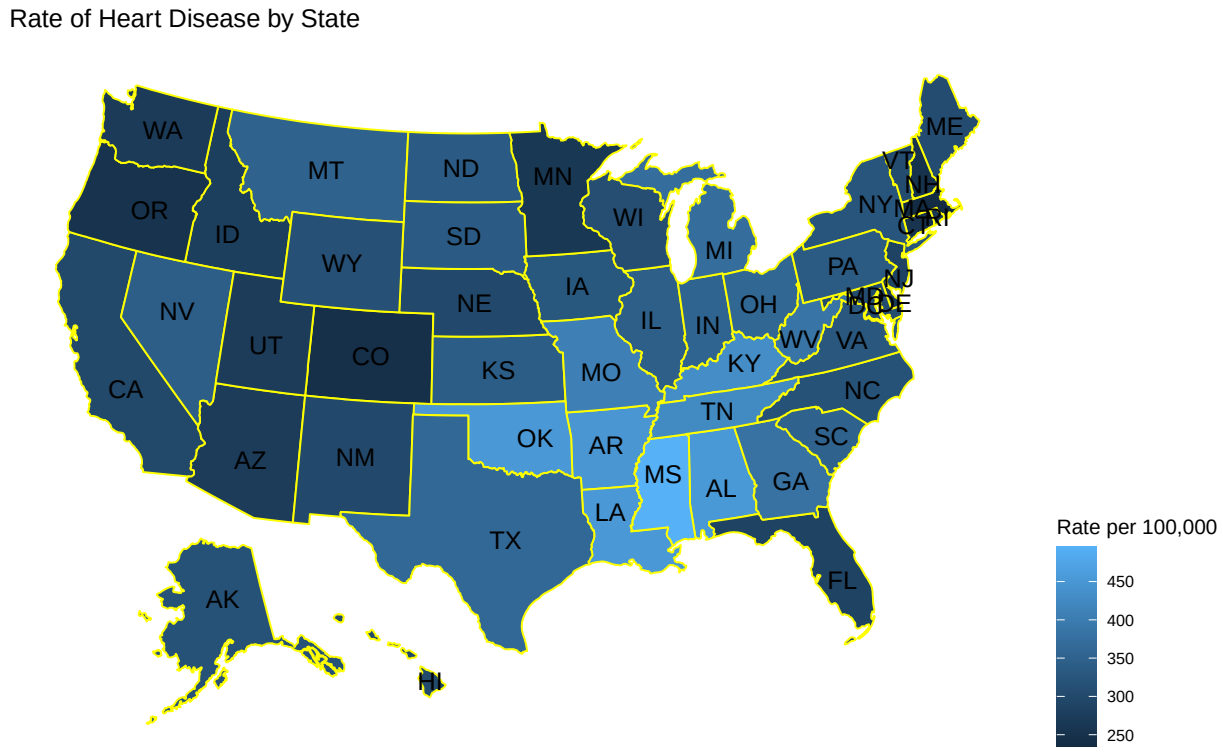
Hypothesis 1

Our first hypothesis is that the rates of heart disease are higher in the southern states of the US. First, we did some research and found scholarly literature that referred to the southern states as the "Stroke Belt" which indicate that the southern states suffer disproportionately from stroke cause by heart disease (Lanska & Kuller, 1995). With this information we can manipulate the data to see if the dataset supports the literature and see if we see higher rates of heart disease in the southern states.

Initially we wanted to see the overall distribution of the heart disease mortality rate across the US by State and that can be seen in the graph below. After we graphed the data, we noticed that there were three US owned territories in our data set: The Virgin Islands, Guam, and Puerto Rico. We did not omit these observations as these territories would not affect the data for the Continental States.

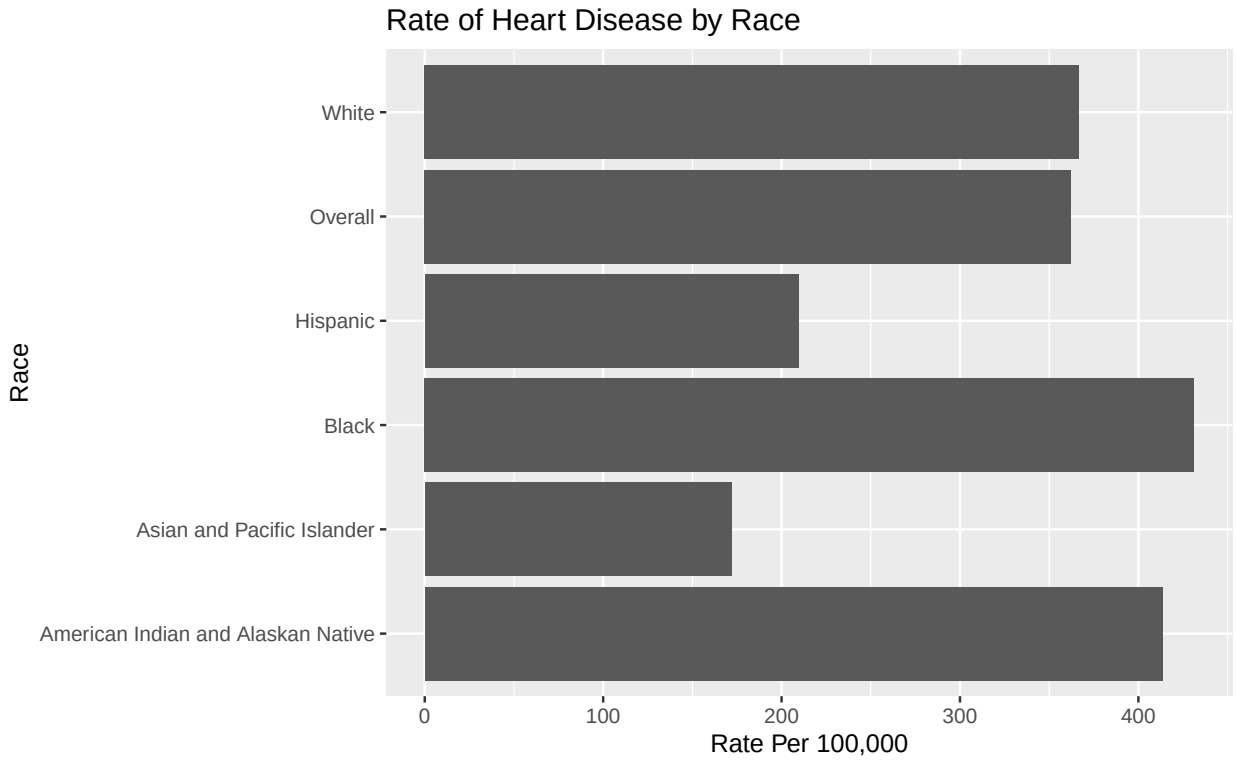


We then graphed the data onto a US map using the 'usmap' package to better communicate the rates of heart disease by region. As noted by the map below, the southern states have higher rates of heart disease than other regions of the US. This map supports both our hypothesis and scholarly literature that rates of heart disease are higher in the southern states. The lighter the blue on the map, the higher the rates of heart disease.



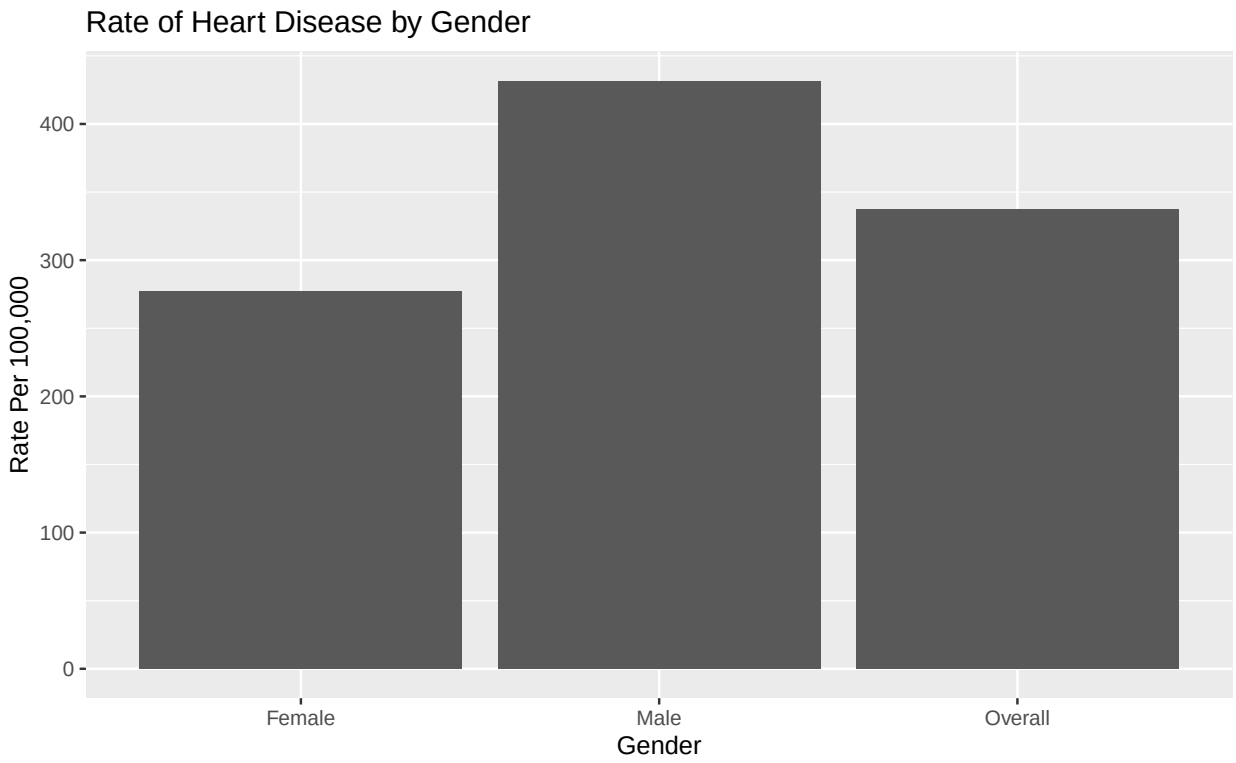
Hypothesis 2

Our second hypothesis is that African Americans have the highest rates of heart disease. After additional research, we found literature to support the idea that African Americans suffer from higher rates of heart disease due to a combination of racial and genetic factors (Yancy, 2006). Since the dataset included a race/ethnicity variable we manipulated the data to separate rates of heart disease by race/ethnicity to see if our dataset supports the literature. After manipulating and plotting the data to a graph, we can see that African Americans have the highest rates of heart disease versus other races. Essentially, the data supports both our hypothesis and the literature.



Hypothesis 3

Our third hypothesis is that men have higher rates of heart disease than women. With further additional research, we found literature that supports the idea that men have higher rates of heart disease than women due to a combination of biological, behavioral and psychosocial factors (Weidner, 2000). The dataset includes a gender variable, so we were able to manipulate the data to show rates of heart disease between men and women. As shown in the graph below, the data supports both our hypothesis and literature that men have higher rates of heart disease than women.



Data-driven Hypotheses

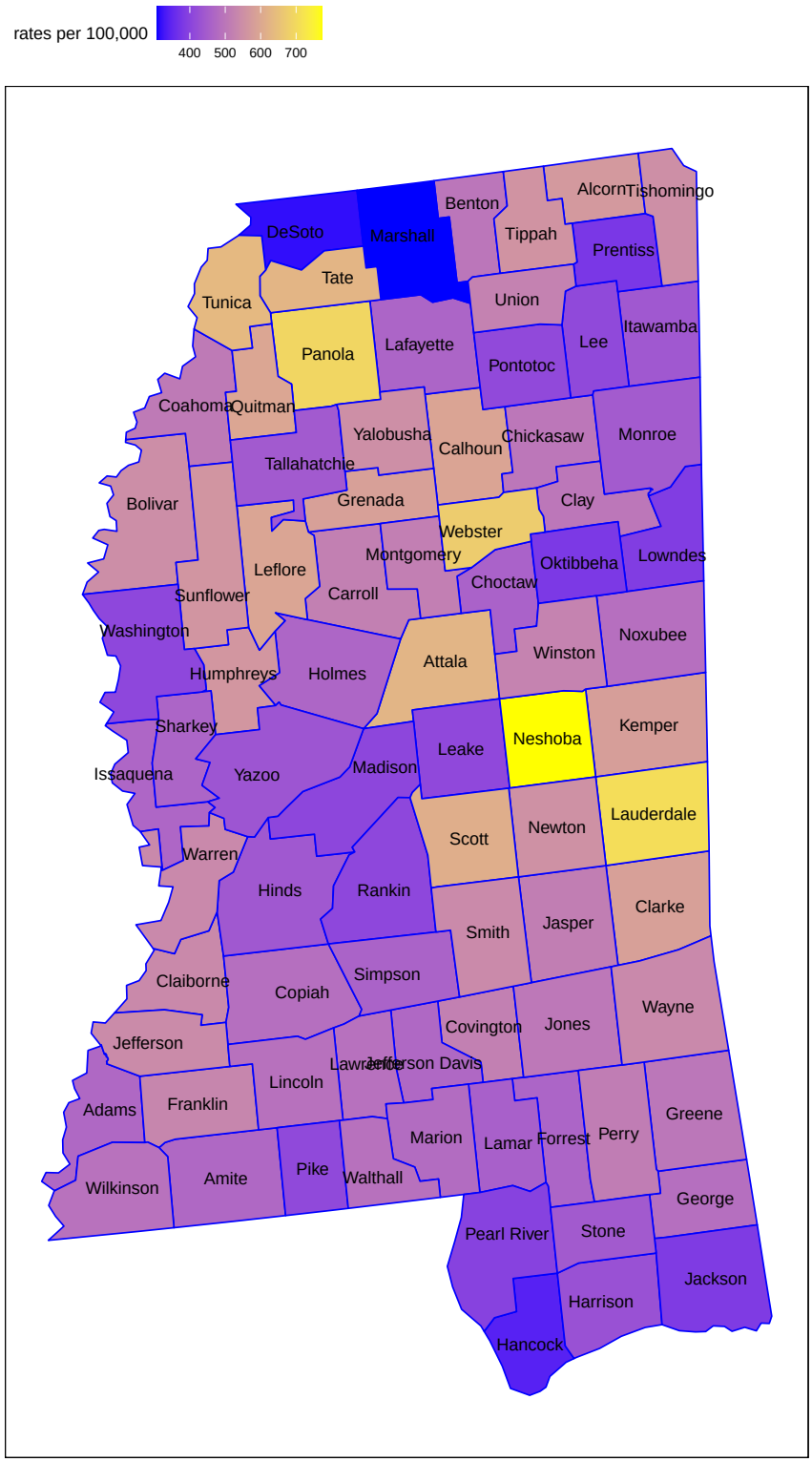
Since our dataset included a county variable, and we utilized the 'usmap' package to graph data onto a map, we were interested in seeing if rates of heart disease were higher in urban or metropolitan areas or in rural areas. We hypothesized that the rates of heart disease are higher in rural areas. In our earlier graphs, Alabama, Mississippi and Oklahoma had the highest rates of heart disease so we chose those three states to support our hypothesis. We also found literature that suggests that rural areas have much higher rates of heart disease than urban areas, and attributed the disparity due to city residents being better educated, better adherence to medication, lower rates of obesity and smoked less frequently (Kulshreshtha et al., 2014).

As noted earlier, we relied on the 'usmaps' package to plot our data on a map but to plot our data by county required FIPS codes instead of the longitude and latitude. We obtained

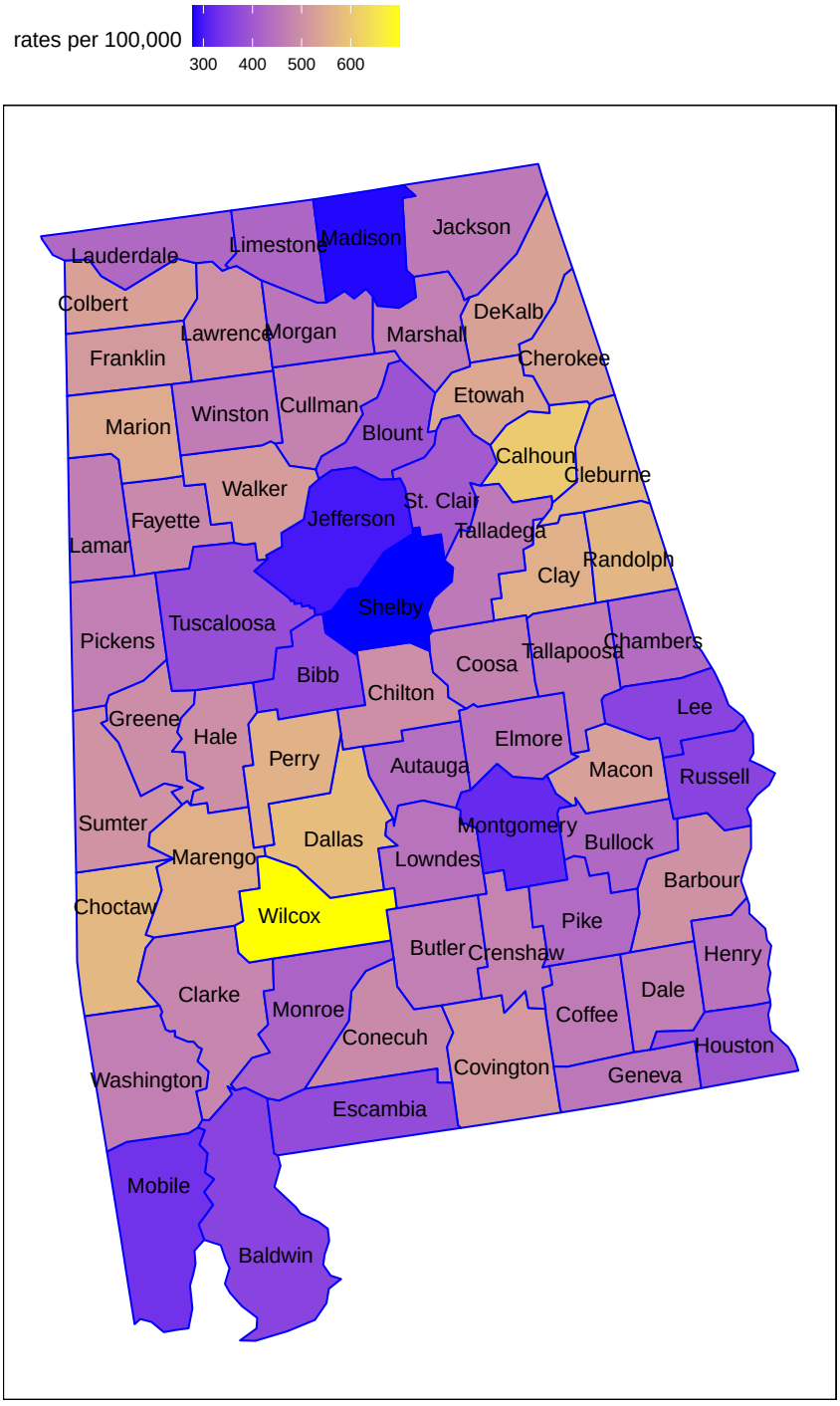
the FIPS codes from the 'Tidycensus' package and matched it with our dataset so that we are able to plot the data onto the maps. Since we are already know the top three states with highest rates of heart disease we first filtered by States, and then merged the FIPS code into our filtered dataset instead of merging the entire FIPS code into the whole dataset. There were numerous NA values in our dataset and we attempted to match the original dataset with the FIPS code but were unsuccessful.

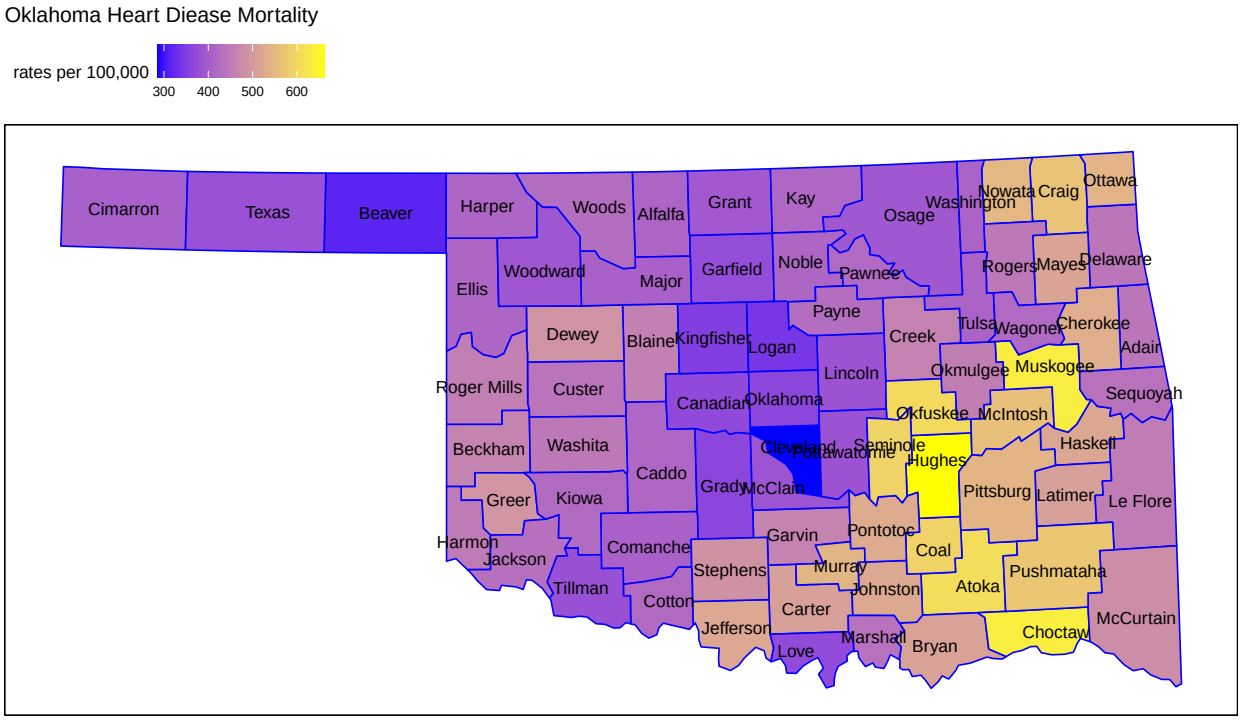
Ultimately, we plotted the dataset, with the FIPS code attached, to the states Alabama, Mississippi and Oklahoma. The results show clusters of areas with high rates of heart disease and low rates of heart disease. We then compared the map to a geographical map of the state to see where the urban and rural areas were and the maps were observed on www.geology.com. The data plotted maps for all three states shown below show that the areas with the lowest rates of heart disease in all three states were at major cities and the highest rates of heart disease for all three states were in rural areas. The maps in addition to the literature supported our hypothesis that the rates of heart disease are lower in urban than rural areas.

Mississippi Heart Diease Mortality

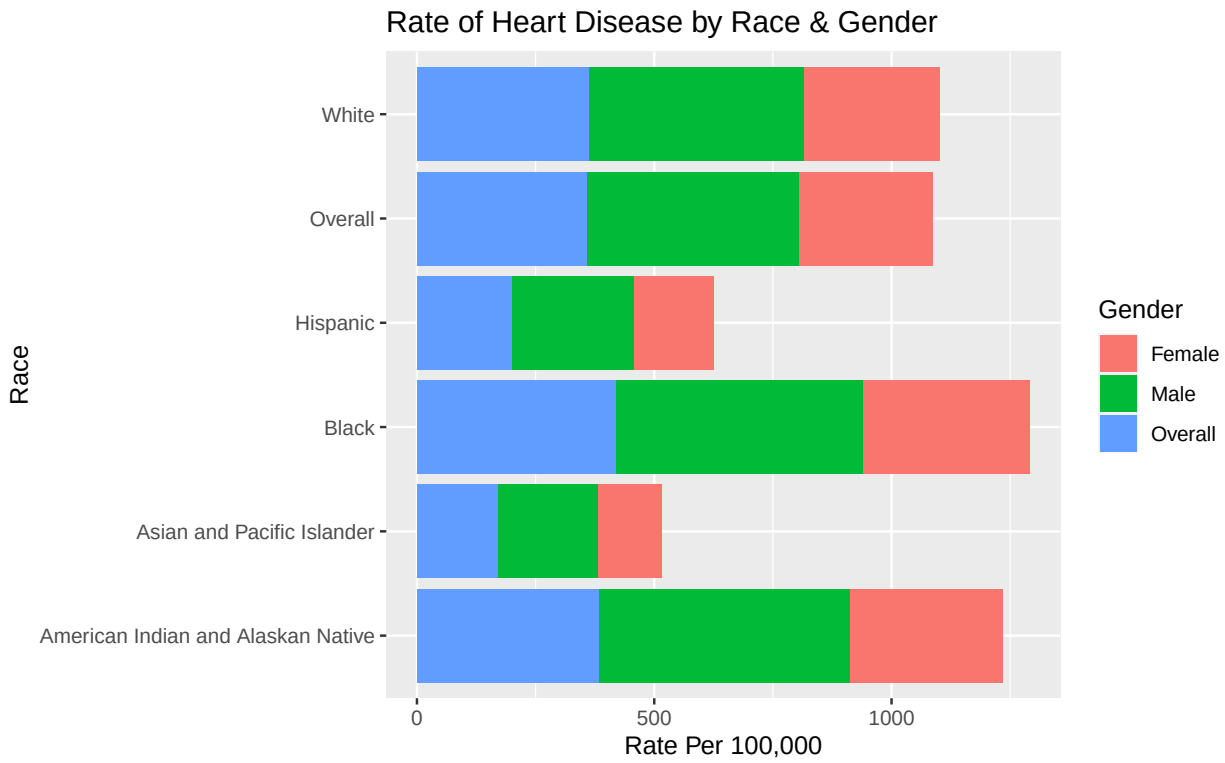


Alabama Heart Diease Mortality





Our analysis so far included rates of heart disease by gender, race and by geographical location. There was one last data driven analysis that we were interested in exploring and that included rates of heart disease in men versus women by race/ethnicity. We hypothesized that the disproportion of heart disease we different in African Americans than other genders. In other words, there were more African American women with heart disease than African American men. We took our existing data where we manipulated the data by race and included the gender variable and the results of the graph are below. The data and graph did not support our hypothesis as the proportion of men and women were the same for all races. We did additional research to see if any literature supported our hypothesis but there wasn't anything to support our hypothesis.



Discussion

The dataset, after certain manipulations, largely supported the literature that we discovered and all three of our hypothesis were supported by the dataset as well. Additionally, there were some hypothesis that we came up with after exploring the data but our data manipulations only supported the hypothesis that heart disease is higher in rural areas was also support by both the dataset and literature. Future research could include additional analysis in different regions of the US. As note in our research, rates of heart disease are different by the various regions of the US. Our research was mainly centered around the southern states. Perhaps, our three hypothesis in addition to our data driven hypothesis isn't supported in all the other regions or states.

As noted in our data driven analysis, the dataset required FIPS codes to be able to graph the data by county in the usmap package. At the conclusion of our research we conducted code

debugging and cleanup and discovered that the original dataset already included a FIPS code column. In the original dataset the column "LocationID" is the FIPS code. We did not realize the original dataset had FIPS codes as we did not know what the LocationID column indicated. For future research, careful examination of the original dataset and determining what each column indicates would be necessary to prevent additional unnecessary work.

References Cited

Kulshreshtha A, Goyal A, Dabhadkar K, Veledar E, Vaccarino V. (2014) Urban-rural differences in coronary heart disease mortality in the United States: 1999-2009. *Public Health Rep*, *129*(1), 19–29. https://doi.org/10.1177/003335491412900105

Lanska D, Kuller L, (1995) The Geopgraphy of Stroke Mortality in the United States and the Concept of a Stroke Belt. *Stroke*, *26*(7), 1145-1149. https://doi.org/10.1161/01.STR.26.7.1145

Weidner G. (2000). Why do men get more heart disease than women? An international perspective.*Journal of American college health : J of ACH*,*48*(6), 291–294. https://doi.org/10.1080/07448480009596270

Yancy, C. (2005). Heart Failure in African Americans. *The American Journal of Cardiology*, *96*(7), 3–12. https://doi.org/10.1016/j.amjcard.2005.07.028