Project_Progess_Report

Richardson, Trevor K./ Liu, Chi-Yun

2020-11-10

Hypothesis

Hypothesis 1: Heart disease is higher in the "Southern States" of the United States
Hypothesis 2: Heart disease is higher amongst African Americans versus all other race/ethnicity

Hypothesis 3: Heart disease is higher amongst males versus female

- The scholarly journal linked here: <https://www.ajconline.org/article/S0002-9149(05)01211-7/fulltext> shows that African Americans are a unique population in regards to heart disease and suffer disproportionately when compared to other populations.
- The scholarly journal linked here: <https://www.sciencedirect.com/science/article/abs/pii/S1047279707001536> shows that African Americans in southern states suffer from stroke due to heart disease versus African Americans in other regions of the US. The article terms the southern states as the "stroke belt" indicating the stroke is more prevalent in the southern states.

In the following, we will start to tidy and explore our data to find some evidence to support our hypothesis.

Load Data and Data Input

We are loading the dataset from: <https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/i2vk-mgdh>

There are 59076 rows and 19 columns in the dataset.

Explore the Data

Here we will summarize and explore the data.

```
#first few row
head(heart_disease)
# structure of data
str(heart_disease)
#summary
summary(heart_disease)
```

```
heart_disease %>%
  summarize(across(everything(), ~sum(is.na(.)))) # check the NA's value, remove NA's
```

Tidy Data

When looking at the heart disease mortality dataset, it is a bit messy and we need to tidy the data to make it look cleaner.

As for the rates of heart disease, there are two columns of note. "Data_value" and "Data_Value_Unit". When using the unique function on Data_Value_Unit there is only one value "Per 100,000 population." We can rename "Data_value" to "RatePer100Thous"

(Rate per 100,000 population), and then remove (not select) "Data_Value_Unit" from our new dataset. After reviewing each variable, we finally keep 9 columns of variables in our new dataset.

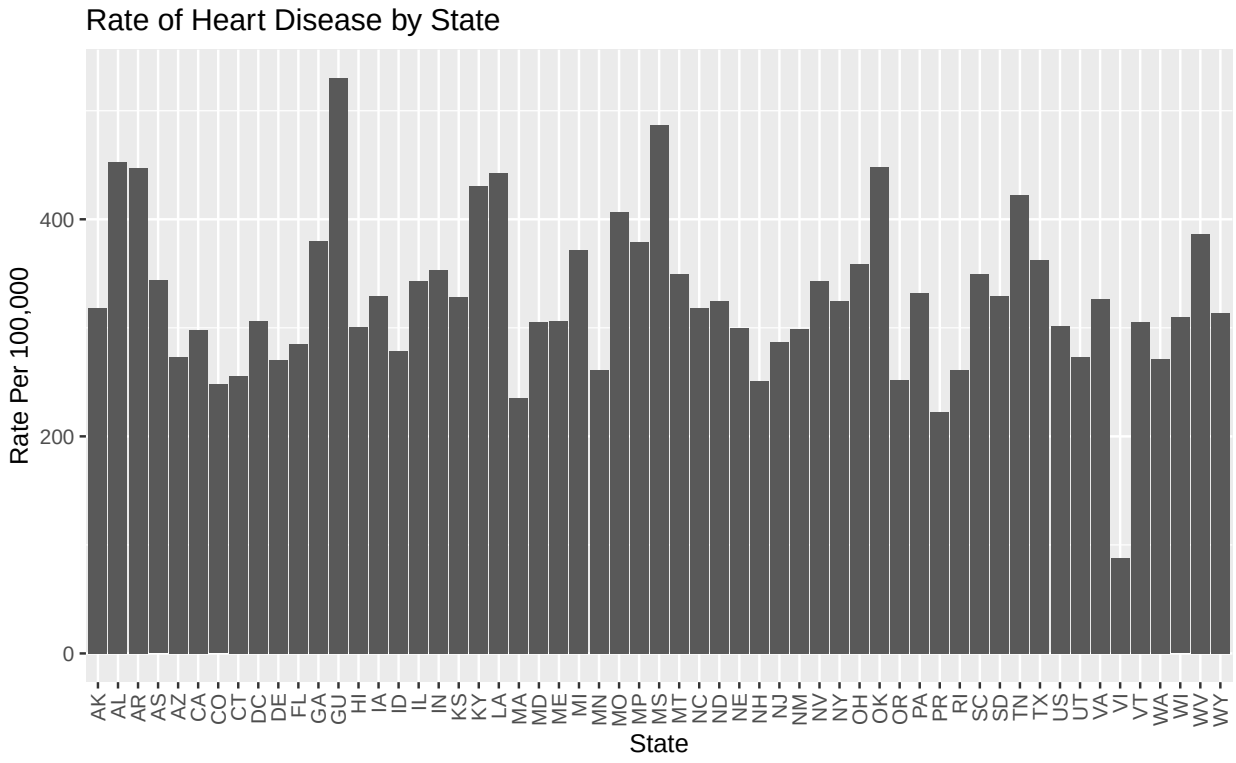
Evidence for Hypothesis

Hypothesis 1

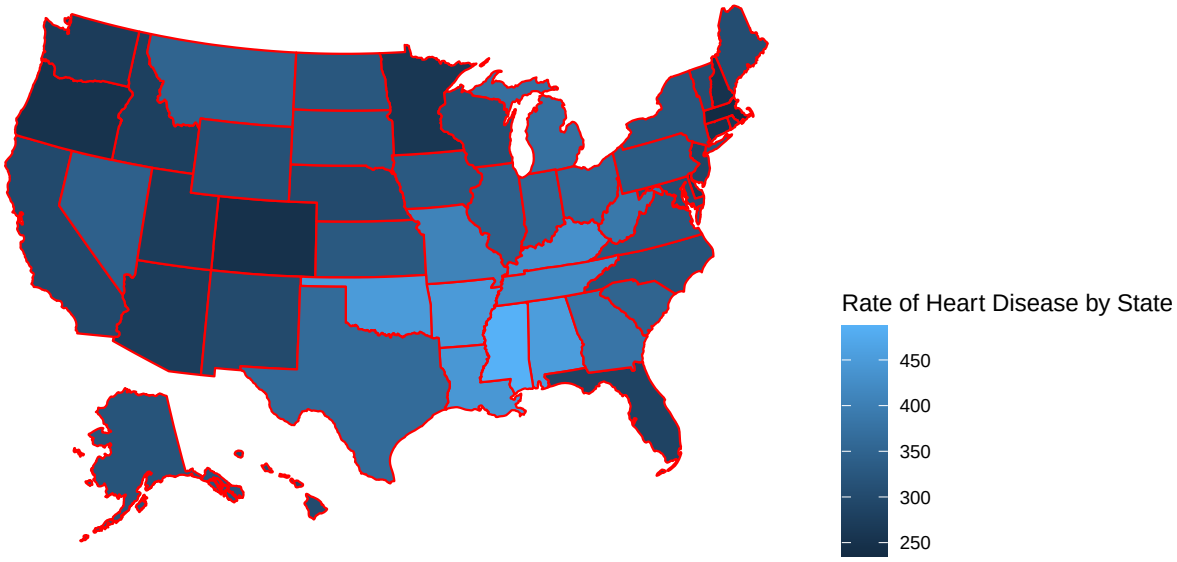
Our first hypothesis is that **heart disease is higher in the Southern States**, this is largely supported by data that show that rates of heart disease are typically higher in the Southern States. As we can see in the US map below, the light blue states is the Southern states and it shows the rates are higher.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Since a graph is a bit difficult to interpret the data since there are 50 states graphing on a US map would be better to interpret the data.



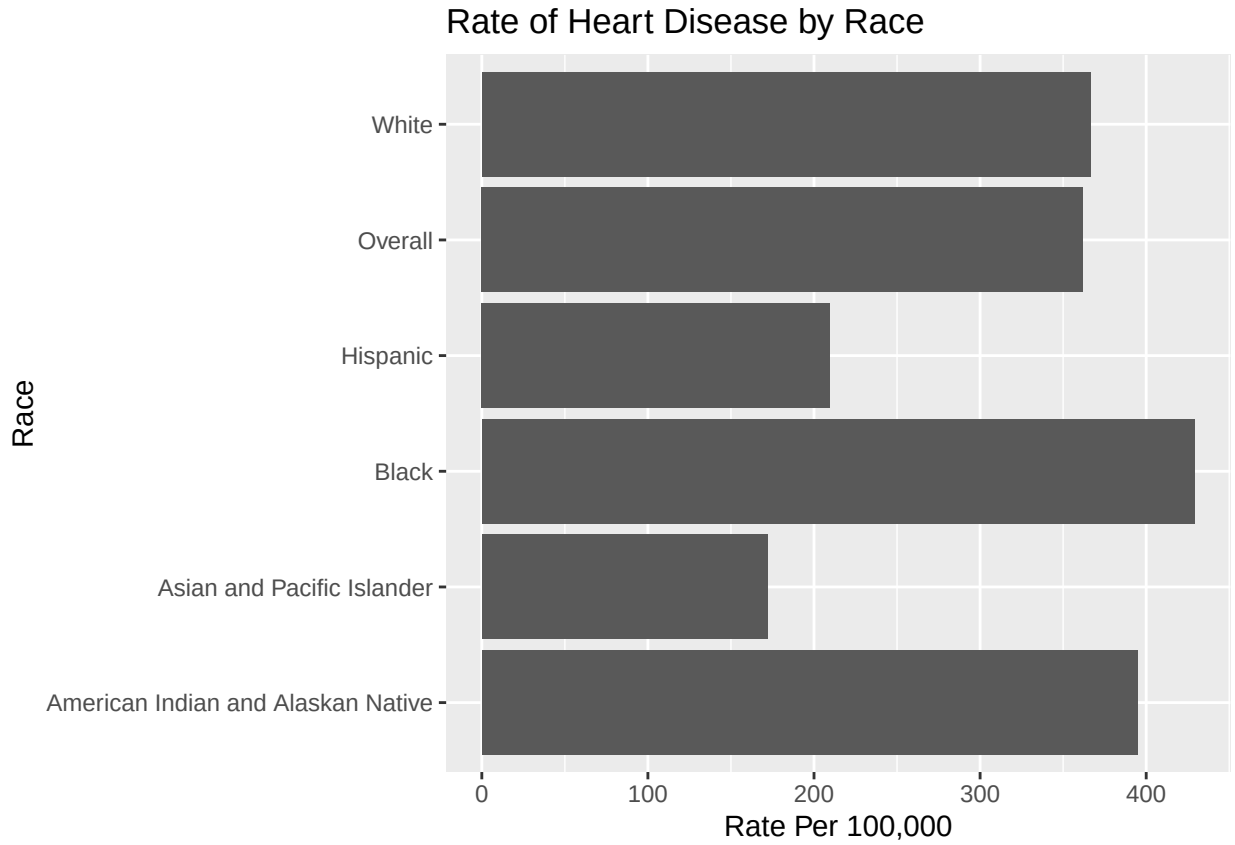
In the map, the light blue states is the Southern states and it shows the rates are higher.



Hypothesis 2

The second hypothesis of us is that **heart disease is higher amongst African Americans**. Based on the plot, we can see that the rate in black is the highest amongst all other race/ethnicity.

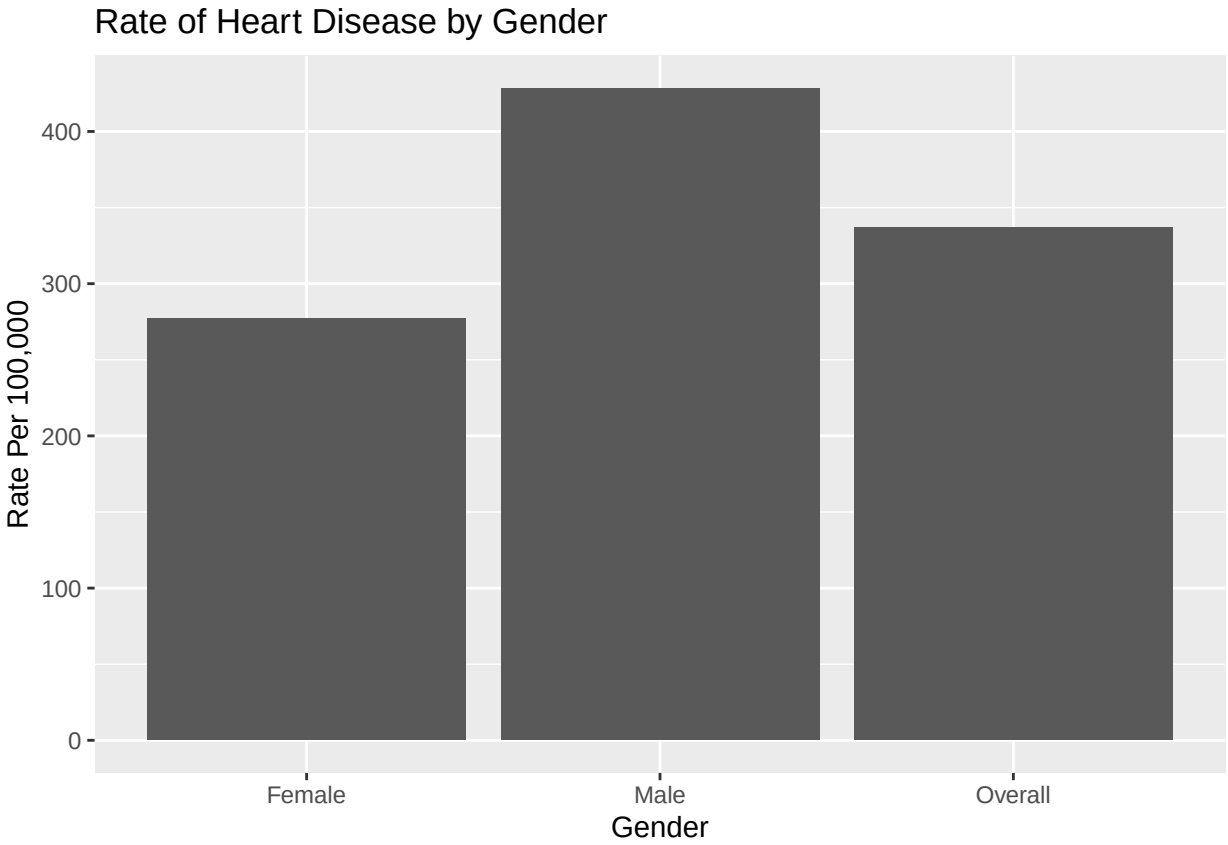
```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Hypothesis 3

The third hypothesis is that **heart disease is higher amongst males**. Based on the plot, we can see that the rate of male is higher than female.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Next Steps

The work we've done so far is that we tidied and explored the data. Through initial analysis, we found the evidence to support our three hypothesis along with two scholarly articles. We also provided a visualization of the evidence in the form of graphs. Our next step is to fine tune the data even further and to find a few more scholarly articles to support our multiple hypothesis, as we are missing an article to support our gender hypothesis. Additionally, since we only have generalized data, for example, we have the rates of heart disease per state but we can further break it down by county and see if its consistent between all states. Essentially, we need more visualizations. Lastly, the final step would be to put all of our research into a presentation that is both visually informative and communicates well.